面向社会化媒体大数据的社会计算

Social Media Big Data

梁循　杨小平　周小平　张海燕　　编著

清华大学出版社

面向社会化媒体大数据的社会计算

梁　循　杨小平　周小平　张海燕　编著

清华大学出版社
北　京

内 容 简 介

本书综合了大量国内外的最新资料和作者的研究成果，介绍了社会计算的定义和研究内容，以社会化媒体大数据为例讨论了数据获取和知识表示，从社会化媒体的网络结构和内容的角度研究了社区发现算法和兴趣社区划分方法，讨论了社会化媒体网络信息的传播问题、跨平台挖掘以及群体智慧的一些相关研究成果。

全书围绕着三个层次展开叙述：数据层（第1～2章）研究社会化媒体以及社会化媒体的数据获取和知识表示；模型层（第3～5章）重点分析了社区发现和社会建模与分析，社区发现是进行社区建模和分析的基础；应用层（第6～8章）研究社会媒体文本挖掘的情感分析、金融决策分析、跨平台的知识发现、群体智慧方面的应用。全书提供了大量的研究算法和应用实例，每章后均附有思考题。

本书的读者可以是对社会计算感兴趣的专业人士，或是对社会化媒体挖掘感兴趣的商业界人士，也可作为计算机应用方向的教材或参考书。

图书在版编目（CIP）数据

面向社会化媒体大数据的社会计算/梁循等编著. —北京：清华大学出版社，2014
ISBN 978-7-302-37456-5

Ⅰ. ①面…　Ⅱ. ①梁…　Ⅲ. ①数据收集－技术　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 171629 号

责任编辑：刘向威
封面设计：
责任校对：梁　毅
责任印制：

出版发行：清华大学出版社
　　网　　址：http://www.tup.com.cn，http://www.wqbook.com
　　地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
　　社 总 机：010-62770175　　**邮　　购**：010-62786544
　　投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
　　质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
　　课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：170mm×230mm　　**印　　张**：8.75　　**字　　数**：174千字
版　　次：2014年12月第1版　　**印　　次**：2014年12月第1次印刷
印　　数：1～　000
定　　价：　.00元

产品编号：059923-01

FOREWORD 

前言

本书面向社会化媒体大数据，介绍了社会化媒体计算的一些典型算法及应用。本书结合现阶段社会计算的研究方向和热点，从社会计算的实验数据源出发，在介绍社会化媒体的基础上，讨论了社会化媒体的集成技术，并根据社会化媒体的特点，将现有的研究分为基于社会化媒体网络结构和基于社会化媒体内容两类，依次分别阐述各相关研究。第1章为绪言，着重介绍社会计算的定义和研究内容。第2章介绍社会计算的实验资料来源，以社会化媒体为例着重介绍其数据获取和知识表示。第3章研究了基于网络结构的社区发现。第4章讨论了基于内容的兴趣社区发现。第5章主要研究了社会化媒体网络中信息的传播分析。第6章讨论了社会化媒体计算应用。第7章研究了社会化媒体跨平台挖掘问题。从更广泛的角度来说，利用群体智慧的算法实际上是社会计算的特殊形式，即通过社会个体的合作或竞争完成某一特定的任务，所以第8章我们也介绍了群体智慧的相关研究成果。

本书的工作受到了中国人民大学科学研究基金项目(10XNI029)的支持。作者的学生也参加了本书的编写，这些同学是朱浩然、林航、申华、施晓菁、周晨曦、纪阳、李启东、李晓菲、李亚平、柴若琪、郑艺、马超、王宇珩等。

由于作者水平和时间的限制，书中一定存在不少缺点和错误，恳请读者批评指正。

编　者

于2014年5月

目录

CONTENTS

第1章

绪　言

本章学习目标

- 理解社会计算的概念
- 了解社会计算的研究内容

近十年来，随着互联网技术的发展，尤其是Web 2.0的兴起，互联网涌现出大量具有交互功能的Web应用程序和网站。它们吸引了大量的用户，并贡献出海量的社会行为数据。与此同时，移动设备的普及和移动互联网的兴起，人们所能获取的社会行为数据将越来越多。因此，社会计算将很可能成为继物理计算和生物计算之后科学计算的新热点，并催生出新的研究领域和方向。

1.1　社会计算定义

社会计算作为一个新兴的跨学科的研究领域，目前还没有一个公认的定义。一般认为，社会计算是一门现代计算技术与社会科学的交叉学科。不过，我们可以从社会计算出现的背景去剖析概念，将社会计算简单概括为“用社会化方法计算社会”，具体包含两层意思，即“为社会计算”和“用社会化方法计算”。国内学者在深刻思考互联网的飞速发展和网络社会化趋势的基础上，提出了社会计算是面向社会活动、社会结构、社会过程、社会组织及其有关系统、社会功能和传播效能的计算理论和方法。

社会计算反映了社会计算研究与服务的对象是社会，包括虚拟网络（虚拟社区）和现实社会，以及从中抽象出来的人工社会。从这个角度来说，通过信息技术方法对虚拟网络进行分析，了解社会已经发生、正在发生、将要发生的事情，准确地

把握社会的动态特征和运行规律，预测政策实施的可行性，为虚拟网络社会的科学管理和政府决策提供参考。

社会计算作为计算科学和社会科学的交叉学科，已成为人们分析、管理和控制社会系统中相关问题的强有力方法。通常情况下，可以从计算科学和社会科学两个方面对社会计算进行认识。从计算科学的角度看，社会计算是研究计算机以及信息技术在社会中的应用，进而影响传统的社会行为的过程；该角度注重微观和技术的层面，并具有较长的研究历史。从社会科学的角度看，社会计算是基于社会科学知识、理论和方法学，借助现代计算技术，来帮助人类认识和研究社会科学的各种问题，提升人类社会活动的效益和水平；该角度从宏观的层面来观察社会，它注重社会知识在现代计算机技术中的应用，并以此解决传统社会科学研究中使用经验方法和数学方程式等手段难于解决的问题。

显然，社会计算的研究对象是社会，它包括现实的物理社会和虚拟的网络社会。从广义来讲，整个 Internet 就是一个虚拟网络，但从狭义来讲，虚拟网络主要指基于 Web 2.0 的，强调以用户为中心的虚拟社区，如 Facebook、Twitter 等虚拟网络。无论是 Web 2.0 还是 Facebook、Twitter 等虚拟社会网络系统，其最大的特点就是强调用户与用户间的交互，实现的是人与人的互联。如何促进人与人的交互是社会计算研究的另一重要内容。随着 Web 2.0 理念的深入，交互的重点已经从传统的人-机交互（Human Computer Interface，HCI）转化为人-人交互（Human Human Interface，HHI）。对不同的应用领域，人人交互的模式不同。例如在微博中，交互方式包括跟帖、回复、粉丝等；在人际关系网中，人人交互一般显性表现为加某某为好友。目前有少数学者从信息系统行为角度对社会网络信息交互模式、基于 Web 2.0 的信息生成模式、Web 2.0 环境下知识共享问题进行研究。

社会计算讲究的是用户协同。随着大量社会网络的产生，以 Web 2.0 思想为核心的社会协同计算模式正逐步应用到诸多领域，如个性化推荐、电子商务、网络营销等。社会计算是一种以“草根”用户为中心、并依靠“草根”用户的用户化方法，一种协同和群体智能的方法，是一种从个体到整体，从微观到宏观的思维模式。许多事件都是由无数网民微不足道的微观行为最终发展成一个重大的社会事件或浩大的工程。从这个角度来讲，社会计算是一种群体智能的计算模式。

虽然社会计算近年来才引起国内外学者的高度重视，但从计算科学的角度来看，社会计算的研究已经有较长的研究历史。随着计算机网络的出现，各类交互软件和 Web 应用程序的出现，使得计算机成为一种新兴的通信工具，它拉近了人们之间的距离，并使得分布在世界各地的用户之间拥有了新的合作和交流方式。因此，从技术的角度来看，社会计算的功能之一就是研究使用计算机技术，构建社会软件（Social Software），为人们的沟通、协作创造一个便利的“虚”环境。基于这个

理解，1994 年，Schuler 就提出了社会计算（Social Computing）的概念。从该层面上看，社会计算是指支持任何社会行为的计算机系统。它通过软件和信息等技术，构建或重构社会环境及社会对话方式。因此，电子邮件系统、论坛、博客、即时通信软件、社会网络服务、Wiki、社会书签以及其他各种形式的社会软件都属于社会计算的范畴。随着社会软件的发展，越来越多的用户参与社会软件的活动，进而产生越来越多的社会行为数据。于是，有学者发现，一方面，通过大量用户的参与，集合群体智慧，可以解决传统方法所无法解决的问题；另一方面，人们还发现利用用户所贡献的海量数据，可以分析、解决许许多多的社会学问题。基于这些新认识，2005 年，James Surowiecki 在 *The Wisdom of Crowds* 一书中，提出了社会计算是利用群体智慧进行计算的概念。其内容包括协同、市场预测、信誉评估、在线拍卖等。近年来热门的“众包”概念就是强调借助计算机网络，有效利用广大用户群体的智慧，解决相关应用问题，属于社会计算的范畴。

与此同时，各类恐怖事件的发生，包括美国的“9・11”、西班牙的“3・11”和英国的“7・7”等事件，促进了人们对社会计算的研究。人们越来越意识到：人们需要构建一种新的信息处理方法，充分挖掘海量社会行为数据，从而获取更多有效的情报内容，进而保障社会公共安全。在这个背景下，2003 年，美国政府提出了“情报与安全信息学”的概念。毫无疑问，“情报与安全信息学”只是将现代计算机技术应用于社会问题解决的一个具体例子。随着信息技术的发展，海量社会行为数据将催生现有的大型计算方法和应用逐步扩展到社会计算的各个领域，进而解决各类社会问题。所有这些都属于计算层面社会计算的范畴。

社会计算的发展离不开社会软件的兴盛、Web 3.0 技术的发展、社会网络分析（Social Network Analysis，SNA）在学术界的持续走热、开源软件的兴起以及人们对这些技术所带来积极影响的信心。以往，人们往往从研究机构、企业、媒体、宗教和政治团体等获取信息；在社会计算的影响下，随着终端设备的普及、信息内容的结构化和获取便捷性以及计算资源的共享，人们的沟通和交流越来越便利，人们越来越愿意从其他个体获取信息。这种根本性的变化无疑将对经济和政治产生极为深远的影响。

经过近百年的发展，传统社会科学，诸如经济学、社会学等领域都形成了一套严谨的体系结构。然而，与自然科学相比，现有的社会科学体系还远远不够完整。正如社会学鼻祖奥古斯特・孔德对社会学的定义：社会学希望使用一种类似于物理学这样的自然科学的方法与理论，统一所有的人文科学学科，从而建立一门经得起科学规则考验的新的人文学科。因此，从社会学的角度来看，社会计算强调以现代计算技术为工具，应用社会科学理论，研究解决社会问题的方法和手段，进而帮助建立社会科学诸多领域的理论和方法学体系。它涉及社会科学诸多领域的许多

重大问题，它以人及社会为对象，研究其建模、实验及评价方法，进而帮助解决社会经济、政治等领域的诸多难题。

虽然，传统社会科学领域已经建立了一套基于数学的定量研究方法，但是，这些理论方法在解决现实问题时，往往得出截然相反的结论。其重要原因在于这些理论所使用的模型往往忽略了现实生活中的某些因素。现实世界是一个复杂的系统，现实世界中的个体是相互联系、相互作用的，人们可以用简单直观的数学理论描述某个或某类个体在某个时刻的行为，却很难用其描述整个现实世界的经济和社会行为。20 世纪 70 年代，某些研究机构开始注意到人类社会中经济和社会系统的复杂性，并开创了复杂性科学的研究领域。计算机以其强大的计算能力，成为人们研究复杂系统的基本工具。学者们提出了复杂性的相关理论，使用计算机进行复杂系统模拟和仿真，并观察系统的相关行为。在此基础上，随着东欧社会的变革，相关研究人员提出了“人工社会”、“人工科学”等概念，以利于研究信息技术对社会和文化的冲击和影响，进而形成了一系列研究复杂性科学的方法。

导致人们更加重视这种宏观层面上社会计算研究的一个直接推动力是美国的“9·11”恐怖事件。该事件使得人们意识到政府应当充分利用各种信息技术所获取的信息数据，挖掘数据信息，并结合社会变化情况，制定合理、适宜的社会政策。因此，如何合理地建立人类社会模型，用计算机进行模拟、测试并验证社会经济政策的效果，成为整个社会的一个迫切需求。从社会科学角度上看，虽然社会计算在一些领域已经有了一定的研究成果，然而由于社会系统的复杂性，在理论和实践中，仍然有许许多多的问题需要进一步研究和解决。毫无疑问，人工社会的研究成果将会成为社会理论研究和未来政策制定的基础，但其研究将是一个长期的过程。人们仍然需要研究如何有效地结合社会科学和计算科学，并最终建立一套符合科学规则的社会科学体系。

显然，从不同的角度，对社会计算会有不同的理解；甚至从同一角度，随着相关技术的日新月异，对社会计算的定义也不尽相同。因此，要正确认识社会计算，就要深刻认识社会计算的本质。虽然，各种各样的社会软件给我们贡献了海量的社会行为数据，例如 Facebook、Twitter 等都网罗了海量的用户，人们在其上进行的文字、图片甚至视频交流。然而，人们使用这些社会软件进行日常交流是否就意味着其是社会计算？我们认为，这些行为并不能解决现实社会中的相关问题，因此，还不能称作真正的社会计算。社会计算的本质是通过社会群体的力量，提升现有解决问题的能力，包括计算能力、信息整合能力、社会行为与社会模型构建、分析和实验能力等。在此基础上，我们认为社会计算是以社会科学理论为指导，以现代计算科学技术为工具，充分利用社会群体的力量，提升计算能力、信息整合、知识发现、决策支持、社会建模、社会模型分析与实验等方面的能力，进而解决社会科学问

题的理论、方法、手段、技术和计算系统。

1.2　社会计算研究内容

社会计算旨在使用计算科学手段，融合社会的力量，提高计算、信息整合、知识发现、决策支持、社会建模、分析和实验等的能力。围绕社会计算的本质，现阶段，针对社会计算的研究主要有群体智慧、数据集成、数据挖掘、决策支持分析、社会网络建模、社会个体和群体建模和分析等。图 1.1 描述了社会计算研究内容的层次结构图。社会计算的研究内容主要可以分为三个层次：数据层、模型层和应用层。目前，针对社会计算的研究，其主要实验资料来源于社会化媒体。数据层主要研究社会化媒体以及社会化媒体的数据获取和知识表示。由于人们可获取的资料来源越来越多，因此，为了获取更全面的社会网络结构和用户信息，人们还需要对多个社会化媒体进行数据集成。社区发现是进行社会建模和分析的基础，社区发现和社会建模与分析是模型层的核心内容。基于所构建的分析模型，应用层研究群体智慧、知识发现和决策支持。

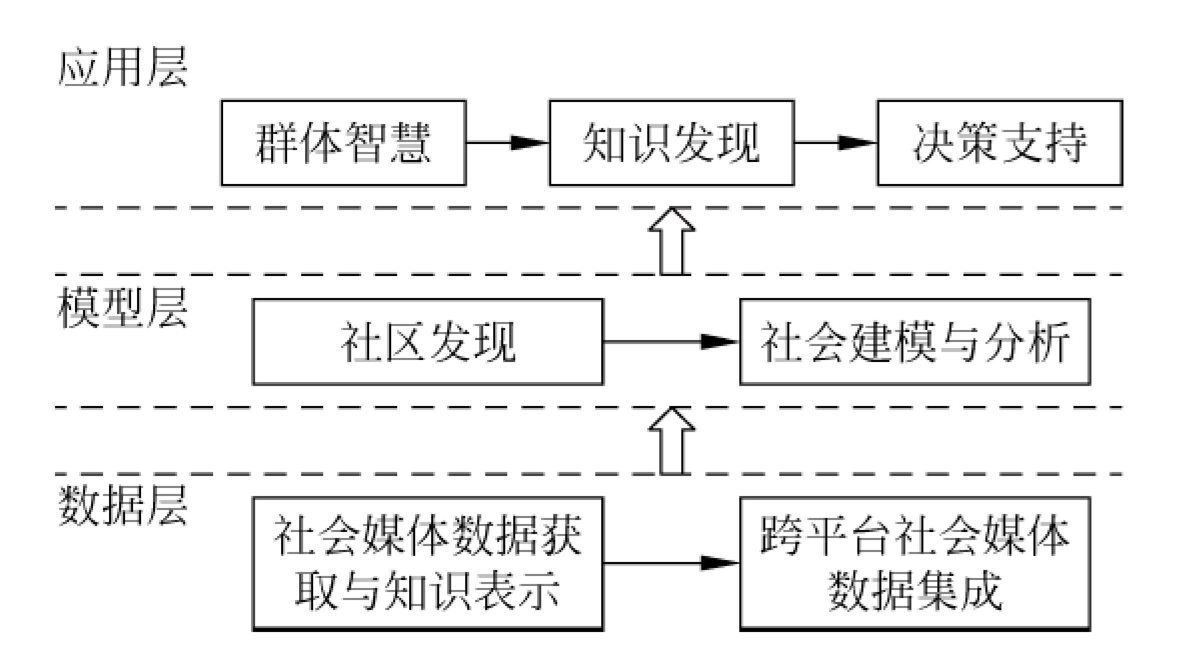


图 1.1　社会计算研究内容的层次结构图

1.2.1　数据集成

随着 Web 2.0 技术的发展和 Web 3.0 概念的提出，互联网上涌现出了许许多多的 Web 应用和网站，并吸引了大量的用户。同时，这些用户又贡献出海量的社会行为数据。而这些数据是现阶段进行社会计算研究的主要资料来源。不同的网站，因其功能不同，用户的使用目的不同，因而，所贡献的数据性质不同。为了能够获取更完整的社会网络结构及社会网络结构中各用户较为完整的数据信息，进而完成更为全面的数据挖掘和知识发现，人们需要对这些网站进行数据集成。

在这些网站中，用户是核心，也是连接这些网站的天然纽带。因此，对这些网

站进行数据集成最有效的方法就是在对用户进行识别的基础上融合网站。

1.2.2 社区发现

社区效应是社会网络中的一种普遍现象。在不同的社会网络中,人们发现个体之间往往存在某些共同的特性,也即社区效应。社区又称组群、簇或模块等,它是指内部之间联系紧密,与外部联系稀疏的一组个体。社区是社会分析中的一个基本概念。从巨大的社会网络中挖掘出社区即是社区发现的过程。社区发现有助于其他社会计算任务的进行和完成,因此,社区发现是社会网络分析的一个基本任务。

社区的发现技术,从最初的图分割方法、W-H 算法、层次聚类法、GN 算法等基本算法,逐渐发展和改进,形成了包括改进 GN 算法、派系过滤算法、局部社区算法和 Web 社区发现方法在内的更具操作性的方法。网络的社区发现可为个性化服务、信息推送等提供基本数据,尤其是在信息时代,社区的存在更加普遍,发现技术应用更加方便,其商业价值和服务价值更大。

1.2.3 群体智慧

当独立的个体合作足够紧密时,其所组成的群体就和单一的机体没有多少差别,并具有更强大的能力。社会是社会个体的集合,它组成了更高的智能,并在计算能力等众多方面超越了个体。因此,社会计算中,群体智慧的研究可能会包含社会科学、计算科学与群众行为的研究。有时,人们又将群体智慧称为集体智慧。

Web 2.0 的交互性,使得人们可以便捷地发布自身的内容。群体智慧建立在现有信息技术和网络化的基础上,可提高现有知识的社会贡献以及知识发现的能力。与个体智慧相比,群体智慧不仅是所有文化在数量上的贡献,它还是所有文化在质量上的贡献。

1.2.4 知识发现与决策支持

知识发现又称数据挖掘。它是从社会化媒体所表示的信息中,根据不同的应用需求,识别出有效的、新颖的、潜在有用的以及最终可理解的模式和知识的非平凡过程。知识发现的目的在于从细节烦琐的社会数据中提炼出有意义的、简洁的知识,从而为决策支持提供依据。知识发现为决策支持提供决策依据,决策支持是知识发现的目的。只有将所发现的知识应用于实际问题的解决,才是社会计算的核心和本质。在社会计算中,知识发现和决策支持的研究包括情感分析、舆情分析、识别与预测、营销服务以及各种个性化服务等。

1.3 本章小结

社会计算是一门新兴的多学科融合的交叉学科,它是现代计算机技术与社会科学紧密结合的产物,为社会科学研究和解决社会问题打开了一扇全新的计算之门。本章首先阐述了社会计算在一般意义下的定义,从不同角度和层面详细分析了对于社会计算的理解;其次,阐述了当前社会计算的主要研究内容,并给出了相关研究内容的层次结构图。

思考题

1. 从一般意义上,谈谈你对社会计算的理解。

2. 与传统的社会科学比较,你认为社会计算在社会问题的研究方法与研究内容会有哪些不同?特点是什么?

3. 目前,社会计算的主要研究内容是什么?谈谈你对这些内容的理解。

第 2 章

社会化媒体及其知识表示

本章学习目标

- 理解社会化媒体的概念
- 熟悉主流的社会化媒体平台
- 了解社会化媒体的数据获取方法
- 熟练掌握社会网络分析软件的使用

2.1 社会化媒体定义

社会化媒体是人们彼此之间用来分享意见、见解、经验和观点的工具和平台，是一种提供给用户极大参与空间的新型媒体。它的形式多样化，文本、图片、视频及语音都可以通过社会化媒体进行传播。早期的维基百科、论坛、博客和近年来发展势头正盛的微博、人人网都是其中的代表。社会化媒体在 Web 2.0 的时代下兴起，区别于以技术为创新点的 Web 1.0 时代，Web 2.0 强调的是技术与用户的结合。社会化媒体中的用户不仅仅是互联网的浏览者，更是信息的制造者。网络用户自发创造、贡献、提取并传播信息的过程构成了社会化媒体的基本生命周期。

社会化媒体具有以下特征。

- 参与：社会化媒体最大的特征就是非常强的网络用户参与性。用户是信息的提供者、评论者，这些信息填充、丰富了社会化媒体的基本框架。
- 公开：大部分的社会化媒体允许、鼓励互联网用户参与发布、评论和分享信息，除了有隐私保护的内容之外，社会化媒体的内容是向用户公开的。
- 交流：社会化媒体中，信息是双向传播的，即实现了媒体与大众的信息交流。

- 社区化：社会化媒体注重集体智慧，用户可以通过现实世界的关系网络或个人兴趣等方式建立、发现属于自己的社区。
- 多样性：社会化媒体的形式不再是单一的文本，图片、视频和语音都可以用来传递信息。
- 多平台：社会化媒体的承载平台具有多样性，网页、计算机及手机客户端等都可以通过接入互联网作为服务平台。

从 Web 1.0 到如今的社会化媒体，互联网提供给其用户的服务越来越接近人类社会的结构与交流方式。社会化媒体中，集体智慧更胜于个人智慧。维基百科就是其中非常典型的代表，3200 万登记用户为完善这一百科全书编辑总数超过 12 亿次。微博也是集体智慧的体现，一个拥有 100 万粉丝的账号在一瞬间就可以同时向 100 万人传递消息，因此微博中的每一个见闻、每一条消息都可能以裂变式的速度传播。人人网的构建所依赖的正是群体关系，人们将现实社会中的人际关系和活动转移到互联网中，实现了互联网从虚拟走向现实的跨越。社会化媒体的多样性与交互性，使它成为互联网用户的宠儿，人们已经将自己社交生活的一部分交给了社会化媒体。

不难看出，社会化媒体在互联网中占据了非常重要的位置。同时，对社会化媒体的研究也是非常有必要的。一方面，社会化媒体对维持互联网的信息传播、网络的稳定及其他属性是至关重要的，另一方面，社会化媒体的分析技术可应用于互联网以及现实社会的拓展中。

那么，应该如何着手研究社会化媒体呢？正如前文所说，社会化媒体强调的是技术与用户服务的结合。同时，用户服务在社会化媒体中更多地体现出用户之间的社会结构、社会活动等特点。因此，对社会化媒体的研究可以从技术与社会科学入手，挖掘两者之间的关系、特点等问题。社会计算正是将现代计算技术与社会科学相融合的交叉学科。社会计算是面向社会活动、社会过程和社会结构等的计算理论和方法。对于社会化媒体而言，网络用户不仅仅是受众群体，更是信息的来源。通过社会计算的方法与技术，可以运用数据深入分析网络用户的行为，探索网民的关注点，提取用户的社交关系网络及挖掘潜在的可利用信息等。

2.2　社会化媒体分类

社会化媒体既是一种服务、一种工具，也是一种媒体、一种平台。

立足于使用者的个人角度，社会化媒体作为一种工具为用户提供了与朋友交流（社交网站）、实时跟进新闻事件进展（微博）、获取参考知识（维基百科）、满足音像娱乐需求（优酷、虾米）、聚集不同渠道的信息（RSS）等服务。

从另一个角度来看,社会化媒体的媒体属性十分明显。在社会化媒体的平台上,无数的个人、非个人信息经由网络中结点(人)的不断过滤和传播,迅速在网络中传播扩散,每则信息可能被传播的范围与其所蕴涵的价值成正比,引起不同的社会反响。在某些热点领域,甚至能产生超越传统媒体的影响力。无论如何,社会化媒体都具有一个共同的特点:内容、消息和知识的消费者也是相应的生产者。现有的社会化媒体从其表现形式、运作模式、采用的技术以及其对自身的定位等方面可以明显地分为几个类别。

2.2.1 博客

博客又称为网络日志、部落格等,是一种由个人或组织管理、不定期张贴新的文章的网站。博主(blogger)通常专注于特定领域,多发布较为完整、主题明确的原创文字。

与其他社会化媒体相比较,博客具备的特征包括:拥有明确的个人领域特征,博客中显示的均为与博主相关的内容;实时性较弱,不会有信息流的表现形式;允许交互,但交互功能并非其核心价值;创作的内容比较完整,通常聚焦于博主日常关注的领域,且内容多属可沉淀的内容,不易失效;通常包含大量有价值信息,具有较强的参考价值。当前博客领域并未出现一家独大或寡头垄断的市场格局,在常见的新浪博客、网易博客等博客平台外,更多精品博客以个人博客的形式出现,使用由个人维护的独立域名而并不依附于平台之上。

2.2.2 社交网络

社交网络全称 Social Networking Service,即社会网络服务,一种可以提供多种交流、交互渠道的互联网应用服务,旨在帮助用户在网络中建立并维护社交关系。而更多时候 SNS 是指 Social Network Site,指基于用户的关系网络并为其提供 SNS 服务的平台网站。在这个网站中,用户因朋友关系或是兴趣聚合成一个个社区,信息在社区内流转并引起共鸣。

社交网络最明显的特征在于用户的社区化,用户通过好友关系或是共同的兴趣形成诸多重叠的动态社区,五花八门的信息在社区内进行流转和扩散。社区化带来的优势在于,联系紧密的社区内的用户间拥有较高的信任度,信息在社区内可以做到快速流转,且被用户认真阅读并获得信任。但与此同时,社区内的紧密度与社区间壁垒的坚固度是呈反比关系的,紧密的社区网络代表社区内的用户结构较为稳定,并没有频繁的新用户加入、老用户撤出,而多个社区间的重叠程度是远小于整个平台中的社区规模的。这意味着信息极易被社区分割成一个个孤岛,在整个网络中,用户看到的内容只能是由好友推荐来的小部分信息,信息往往很难实现

跨多个社区的流转。

在结构上，社交网络是一种旋涡式内敛的结构，群聚效应非常明显。一个现实的圈子中使用该平台的人越多，就会继续吸引更多的人进入这个平台，而活跃用户的增长会继续增加用户对平台的黏度；但一旦平台上的用户开始流失并到一定程度后，小圈子的结构会开始崩溃，用户持续流失甚至令此社区消失。因此用户间的交互是平台的核心价值所在，只有用户间的频繁交互才能令用户以更高的热情持续使用该平台，增强用户黏度，避免用户流失。

社交网络的具体产品繁多，具体可以分为两个类别。一是基于好友关系的强关系社交网站，以 Facebook、人人网、开心网为代表。在这类网站中，社区结构借助用户间的好友关系形成，信息借助好友的推介进行流转。另一类是基于内容的弱关系社交网站，以 Flickr、豆瓣、时光网为代表。在这类网站中社区结构是基于对主题、内容的区隔形成的：用户根据自己的爱好聚集在某个主题或内容周围形成社区，而由于主题的限定，其下的信息更是被局限于此区域内，几乎不会发生跨社区的流转。

2.2.3 微博

微博即微博客的简称，是一个立足于用户关系、信息分享、传播以及获取的平台。用户可以建立关注（Follow）关系，以 140 字左右的文字更新信息，并实现即时分享，是社交网络的一种特殊形式。

虽然名字是微博客，但与博客相比，其定位、功能等实际表现明显更接近于社交网站，较为特殊的是，微博中的关系基础是单向的关注机制，用户间因实际的朋友关系互相关注与用户因兴趣对其他用户单向关注的情况共存，使得微博中强弱关系交织在一起，形成了更为复杂的社区结构。除此之外，微博中组织用户大量驻扎且具有相当的活跃度，在个人用户多在发送社会意义极小的生活琐碎信息的同时，组织用户发送严谨的通告类信息，并引发大量个人用户参与讨论，令信息流中具有社会意义的信息比例增长，更具有阅读和研究价值。

2.2.4 分享平台

分享平台是一个概括的称呼，包括 YouTube、优酷等视频分享平台，虾米为代表的音乐分享平台，Flickr、Instagram 等图片分享平台以及 MBALib、百度文库等文档资源分享平台。

此类社会化媒体均专注于某一类型领域，以分享为核心价值，向各自的方向发展，但具有几个共同之处。

首先，分享平台中所涵盖的内容完全覆盖并超出其对应传统媒体的内容。这

有两重的含义，一是传统媒体中的热门资源，会被用户分享到线上，在为平台带来大量流量的同时也反过来帮助传统媒体吸引关注和人气；同时许多由于资源限制而被传统媒体放弃或忽视的长尾资源可以在线上展示，而这正是分享平台明显胜于传统媒体之处。

其次，用户间的互动并不贡献新的内容，而是成为用户活跃的动力组成。其他社会化媒体中用户间的互动同样是整体内容很重要的一部分，但在分享平台中，用户间的互动往往从属于某些信息之下，并且更多的是在扮演一种推动用户分享行为的动力源的角色。在这里，用户间的互动更多的是通过“赞”、“顶”、“踩”这类的方式来进行，并为上传资源的用户带来成就感，鼓励其分享的行为。在单纯的用户互动之外，平台本身往往会推出各种形式的积分、勋章等形式，以鼓励用户进行分享。

另外，分享平台对于数据挖掘、推荐算法等技术的需求更高。在分享平台中，用户间的联系并不紧密，用户获取信息的来源极少来自好友的推介，这种途径的缺失使得为用户推送更适合其口味的内容需要其他有效方法的补充，以提高用户黏度。为了达成这一目标，传统的分类点击排行需要考虑许多新的因素，例如内容的时效性、主题的即时流行程度、不同类型用户的差异化的时间分布、具体内容分类的模糊性等因素；而在点击排行之外，还需要有基于内容相似性的推荐、基于用户相似度的推荐以及其他个性化推荐的方式作为补充。

最后，分享平台打破传统媒体中的诸多限制，用户可以随时随地浏览内容，这也是许多用户以浏览视频网站取代收看电视的习惯的原因。这带来的另一个好处是，分享平台可以以较小的边际成本为用户提供便利的检索渠道。

2.2.5 论坛

论坛全称为 Bulletin Board System 或者 Bulletin Board Service，是一种强交互性的电子信息服务系统，用户可在 BBS 站点上以公开的形式获取、发布信息并与其他用户讨论。论坛可以覆盖的范围极其广泛，既可以有综合型的论坛，也可以有深入度极高的专题论坛。论坛一般由站长创建，并设立各级管理人员协助管理。创办者可依据其创建理念界定论坛的主题、讨论的范围、论坛内的行为规则以及管理人员的具体权限等。

论坛最大的特点在于尽管大量存在，但并不形成稳定的关系网络。具体表现为，当其他社交网站中的用户尽可能地维护当前账户的存在感、可信度以及关系网络时，论坛中并没有成熟的结交和维系朋友关系的功能模块，同时许多用户使用新的账号来避免发言被锁定，以达成在讨论中说出更加符合自己内心的想法而不必有后顾之忧的目的。这是一种优点，让用户可以真正不受拘束地畅所欲言，但同时

也存在弊端,因为可以不必为自己的发言负责,造谣、挑起争端等会破坏论坛甚至是整个网络风气的事情层出不绝,带来网络时代的诚信危机,而论坛中所有信息的真实性和可靠性值得推敲。

2.2.6　知识协作

知识协作是一个可以供多人协同协作的系统。Wiki 站点可以由多人(甚至任何访问者)维护,每个人都可以浏览、创建或更改 Wiki 文本,对共同的主题进行扩展或者探讨,实现快速的信息整合,而 Wiki 的写作者也自然构成了一个社群。

Wiki 最主要的特征在于其开放性。开放性是指社群内的用户可任意创建、修改或删除页面,而这些变化可被任意来访者观察到。开放性使得 Wiki 具有可增长与可汇聚等特征,页面内的任意概念均可通过链接创建新的页面,通过这种方式系统可以不断地增长;而系统内多个内容重复的页面可以被汇聚于其中的某个页面中,并改变相应的链接结构。

Wiki 作为一个群体协作的平台,写作社群内的用户间享有平等的地位,并不因其受欢迎等因素拥有书写内容被优先录用的特权。而在其他社交平台中,往往更具人气的用户其发布的内容被传播的范围更大,从这个方面上来说,Wiki 拥有比其他平台更加平等的特点。由于内容完全由用户创作,所以鼓励用户进行高质量的内容创作是 Wiki 平台的核心所在。通常情况下,平台会建立较为完整的积分、荣誉等奖励体系,激发并依靠用户的成就感来维系其创作热情。

由于是共笔性质,创作、审核均鲜有专业人士,其公信力受到置疑。另外涉及政治、宗教的文章也会因不同国家、政治立场或不同语言用户背景的影响,导致出现编辑战、审查或屏蔽。网络百科的编者往往只是义务参与撰写,并不是该领域的专家,种种因素都使网络百科全书的素质比不上传统百科全书,在较为严肃的时刻并不够可靠。

2.2.7　即时通信

即时通信是一种基于互联网的即时通信服务,国外以 MSN(Microsoft Service Network)为代表,国内则以腾讯 QQ 为代表。即时通信利用互联网线路,通过文字、图片、语音、视频、文件等多种方式进行交流与互动,为用户提供更加便捷和经济的沟通渠道,同时成为人们工作、学习交流的平台。

大部分的即时通信服务提供到场提醒(Presence Awareness)的特性——显示联络人名单、联络人是否在线上以及能否与联络人交谈。各即时通信程序相互独立,无法互通,这使得即时通信软件之间的斗争极为激烈。目前互联网与移动互联

网市场中占有率较大的即时通信软件包括 QQ、MSN、Skype、微信与 WhatsApp。

随着即时通信发展的成熟,即时通信软件已经从纯粹的实时接收和发送信息的应用软件转变为一个集成多种功能的平台、一个互联网入口。在交流的过程中随时选取并搜索聊天中出现的内容,显示交流对象最近的社交动态,根据当前位置和时间推荐商品,在软件保持开启状态的同时推荐并播放音乐,系统会话窗口通知有新的未读邮件,上述已经实现的功能融合均反映出即时通信软件作为互联网和移动互联网入口,集成多种社交媒体的可操作性和良好前景。

2.2.8 垂直社区

垂直社区是针对某特定人群或特定范围内容的社区,它的具体实现形式不受限制,可以是社交网站,如职业社交网站 LinkedIn;可以是变种的论坛,如专注回答的百度知道、专注美食的大众点评;可以是传统的门户网站,如面向球迷的虎扑网。其共同的特点包括两点,一是专注于固定领域,二是用户贡献内容往往具有更高的质量。

用户的需求越来越精细化,但传统的综合性网站中所包含的超大规模数据使得用户对某特定领域的信息需求需要以较大时间成本才能得到满足,同时得到的信息往往质量参差不齐,数目巨大,难以方便地转化为用户掌握的知识。在这样的情况下,用户可能要借助搜索引擎来帮助自己找到需要的信息。但搜索引擎的不足之处在于,当涉及较为具体的问题时,搜索出的答案往往不够细致和充足,因为搜索引擎存在搜索盲区和无法实现全文搜索的局限。此时,垂直社区往往能够满足用户的需求,当用户需要获取一本书的评价时,可以去豆瓣读书进行搜索;需要制作旅游攻略时,可以在蚂蚁网、穷游网等网站中获取更多、更详细的资料和经验。

垂直社区的优点在于,细分的领域意味着用户专注于这一领域进行深度讨论,且用户中往往包含此领域内的专家,因此在垂直社区中,很容易找到专业水准极高的高质量内容。垂直社区的定位使得导购和精准广告这两种商业模式可以非常容易且高效地实现。

2.2.9 搜索引擎

搜索引擎是指自动从互联网上搜集信息,在对信息进行组织和处理后,为用户提供检索服务,将相关的信息展示给用户的系统。搜索引擎包括全文索引、目录索引、元搜索引擎、垂直搜索引擎、集合式搜索引擎、门户搜索引擎与免费链接列表等。而百度和谷歌等是搜索引擎的代表。

新一代的搜索引擎的发展方向是个性化搜索,在猜测用户的真实意图和挖掘

个人偏好的基础上为用户呈现尽可能符合其个人需求的搜索结果。

在个性化的搜索引擎技术中，受到业界广泛关注的是社会化搜索（Social Search）：把用户的社交网络加入搜索引擎中，让用户的朋友来为搜索结果排序。如何借助社交网络的平台和数据开发出基于社会化网络的搜索排序算法，为用户提供更贴合他们需求的搜索结果，Google、雅虎和百度均进行了初步的尝试。而除了利用社会关系来进行个性化搜索外，利用个人信息同样是一个探索的方向。例如利用位置信息预判用户的搜索意图，利用用户发布过的图片、信息、视频、评论、曾经的搜索历史甚至是浏览的行为轨迹等数据，对搜索的真实意图、用户的喜好等进行预判，来提供个性化的搜索结果。

从用户的角度看，新一代的搜索将会是一种连接。将用户与用户连接在一起，提供好友采纳的结果为用户提供决策支持；将用户与产品连接在一起，提供一步到位的购买页面，展示跨平台的商品对比；将用户与信息连接在一起，为用户整理有价值的链接甚至是直接整理信息，而非仅提供非智能的链接入口。

新一代的搜索引擎在传统意义之上向前增加了对用户意图的预判，向后增加了对信息的处理，将搜索与服务直接相连。而在这种个性化的智慧搜索的背后是大数据时代的支撑，只有通过对用户行为的大数据的提炼与分析，才能洞察搜索背后的真实需求，为用户提供更加智能的搜索服务。

虽然不同类型的社会化媒体边界明确，较易区分，但在现在的中国互联网市场中，社会化媒体产品均在集成多种类型，增加用户黏度，抢占市场。典型的代表包括两类。

第一类是以即时通信工具为入口，集成多种社会化媒体产品，使用统一的风格在一个客户端进行展示。例如腾讯QQ，将腾讯公司的其他社会化媒体产品，如朋友网（SNS）、腾讯微博（Micro-Blog）、搜搜（Search）、QQ空间（Blog）、QQ音乐（Sharing）等集成在QQ客户端中。用户可很方便地进行多个产品的浏览和使用。类似的在手机端有微信：即时通信功能、与博客神似的订阅账号、SNS类的朋友圈等聚合在一个客户端中。

另一类则提供接口或增加新功能，而非将产品本身集成在一起。如新浪微博，本身是微博客；由于对字数的限制，用户无法利用新浪微博进行较长的叙述，因此引入了带有博客色彩的长博客——以图片形式展示博客的内容；提供博客的链接，用户可在创作新的博客后，在微博中自动发布带有链接的摘要；多媒体嵌入技术引入视频分享与音乐分享；提供具有论坛神韵的微话题功能，可以一人提出话题，多人在话题下留言参与话题；提供微博平台内的全文搜索，搜索结果排序受用户关系网络影响；所具备的私信功能同样是即时通信的一种。

2.3 主流社会化媒体

表 2.1 列举了部分知名度较高的社会化媒体类别，包括博客、论坛、媒体共享平台、社交网络、知识协作等。

表 2.1 部分知名度较高的社会化媒体类别

博客(Blog)	新浪博客、网易博客、百度空间
社交网络(SNS)	微博、人人网、开心网、QQ 空间、Facebook
微博客(Micro-Blog)	Twitter、新浪微博、腾讯微博、朋友圈、
共享平台(Sharing)	YouTube、优酷、土豆、Flickr、Instagram、虾米、百度文库、MBALib
论坛(BBS)	猫扑、天涯、百度贴吧
知识协作(Wiki)	维基百科、百度百科
即时通信(IM)	QQ、MSN、Link、微信
垂直社区	问答类社区(百度知道、知乎)，职业社区(LinkedIn)
搜索引擎(Search)	Google 个性化搜索、淘宝的搜索

2.3.1 维基百科

2001 年创办至今的维基百科(Wikipedia)是一个自由内容、协同编辑且多语言的网络百科全书，通过 Wiki 技术使得所有人都可以简单地使用网页浏览器修改其中的内容。维基百科一词源自其网站核心技术 Wiki 以及具有百科全书之意的 Encyclopedia，形成了新创造出来的混成词 Wikipedia。网站的目标及宗旨是为全人类提供自由的百科全书。

网站由来自世界各地的志愿者合作编辑而成，总共收录了超过 2200 万个条目，其中又以英语维基百科超过 415 万个条目的数字排名第一。维基百科允许任何访问用户使用网页浏览器自由阅览和修改绝大部标签页面的内容。据统计在维基百科上大约有 35000000 名登记注册用户，其中有 100000 名积极贡献者长期参与编辑工作，整个网站的总编辑次数已超越 12 亿次之多。截至 2013 年 1 月为止维基百科整个计划总共有 285 种各自独立运作的语言版本，且已被普遍认为是规模最大且最为流行的网络工具书，平均每天能够有超过 80 万人次的浏览记录。根据知名的 Alexa Internet 网络流量统计数字指出，全世界总共有近 3.65 亿名民众使用维基百科，且维基百科也是全球浏览人数排名第六高的网站(最高纪录是排名在第五名位置)，同时也是全世界最大的无广告网站。

由于维基百科是基于互联网运行的，因此来自全球各地的贡献者可能在浏览相同语言版本的维基百科时却使用不同的方言，又或者受到不同国家的习惯用语

影响而使得彼此用语出现些微差异。这些差异可能导致条目的文字拼写或者用法习惯上出现冲突(例如英语用户就 color 和 colour 等拼法或者是中文用户的繁简体转换问题等),或者是受到不同地点社会环境的影响下使得对于条目内容的观点不一。另外尽管在各种语言版本的维基百科之中也有如同“中立的观点”般普遍施行的方针,然而许多语言版本的维基百科仍然受到主要使用用户国家的法律限制,这使得各个维基百科计划在方针和做法上并非一致。其中最为明显的例子,则是个维基百科必须依照相关法律限制决定是否能够根据许可而采纳自由内容或者合理使用的内容。

维基百科是个民主制、精英制、独裁制的混合。通常大部分的内容,由一般的维基人讨论、修改,通常为民主的形式。维基百科的系统里同时有资深的维基人担当管理员,负责清除破坏及封锁恶意破坏者的账户。非常敏感的议题则由吉米·威尔士最后把关。

2.3.2 新浪微博

新浪微博是一个由新浪网推出,提供微博客服务的网站。用户可以通过网页、WAP 页面、外部程序和手机短信、彩信等发布 140 汉字(280 字符)以内的信息,并可上传图片和链接视频,实现即时分享。

新浪微博提供的功能主要包括:

(1) 发布功能 用户可以像博客、聊天工具一样发布内容;

(2) 转发功能 用户可以把自己喜欢的内容一键转发到自己的微博上,转发时还可以加上自己的评论;

(3) 关注功能 用户可以对自己喜欢的用户进行关注,成为这个用户的关注者;

(4) 评论功能 用户可以对任何微博进行评论;

(5) 话题功能 用户可以在两个 # 号之间插入某一话题,则发出的微博可很方便地通过话题被搜索到;

(6) 私信功能 用户可以点击私信,给新浪微博上任意的一个开放了私信端口的用户发送私信,这条私信将只被对方看到,实现私密的交流。

新浪微博采用的推广策略是邀请明星和名人加入,开设账号,并对他们进行实名认证,认证后的用户在用户名后会加上一个橙色字母 V,以示与普通用户、微博达人的区别,同时也可避免冒充名人微博的行为。

目前新浪微博上大量的媒体工作者、政府部门、企业公司和民间组织,将其作为一个发布和交流信息的平台,同时它也成为社会话题发生和讨论的重要平台。可以说,现在微博逐渐开始成为一种实时民意调查系统,成为一个舆论监督利器。

2.4 社会化媒体大数据

目前，社会化媒体每时每刻都在产生巨大的数据。例如，Facebook注册用户超过10亿，每月上传的照片超过10亿张，每天生成500TB以上的数据。一般认为，从规模上超过1PB就到了大数据的范畴（1024GB＝1TB，1024TB＝1PB，1024PB＝1EB，1024EB＝1ZB）。大数据的特点是：（1）海量性，即数据的量要大；（2）多样性，显然社会化媒体具备多样性（网络日志、图像、音像、文字、数值、XML、HTML、各类报表）；（3）高速性；（4）大数据的价值密度低，这也导致大数据挖掘需要新的、更快速的方法。在数据分析上，如果在上千台机器上设计挖掘算法，就非常的不方便。如何把数据的潜在价值发挥出来，是一个挑战。显然，社会化媒体数据是大数据。

2.5 社会化媒体大数据获取方法

2.5.1 维基百科数据获取方法

维基百科是一个动态的、允许任何人自由访问和编辑其中的文本及条目的网络百科全书。按照对社会化媒体的分类，维基百科属于群体协作类社会化媒体，其主旨就是允许多人通过群体协作进行知识共享。

维基百科按照一定的规则对词条信息进行分类，“人物”类是其中一大类。维基百科人物类收录了包括重要的历史人物和对时事有关键影响的人的条目。对于人物的收录，维基百科有一定的收录标准，同时，群体协作的编辑方式使收录的信息具有较高的可信度。基于这样的可靠性，可以以维基百科中的人物类为例，通过社会计算的方法对人物类中的关系网络进行挖掘。通过对关系数据的挖掘，可以进一步探索在非强关系的制约下，重要的社会人物之间的关系信息。

截至2012年11月2日，中文维基百科的条目数已经突破60万，全球所有282种语言的独立运作版本共突破2100万个条目。维基百科提供完整的数据库转储文件给感兴趣的使用者，数据库转储文件中的信息以不同的文件格式进行储存，同时保证定时的更新。可以通过维基百科官方网站获取最新的数据库转储文件。2013年12月8日更新的中文版维基百科数据库转储文件解压后的XML文件大小为684.8M，本书将以此版本为例，对维基百科中数据获取的方法进行说明。

中文版维基百科数据库转储文件包含了当前版本的条目、模板、图片描述、基本的元页面。在提取并存储人物类的页面信息前，需要提供一个页面名称列表。由于每个人物在维基百科中对应一个唯一的页面，因此可以将人物名称作为页面

名称列表来存储页面。这种存储方式保证了人物名与维基百科中的词条对应，使之不会出现歧义或找不到对应页面的情况。

维基百科人物类数据获取示意图如图 2.1 所示。

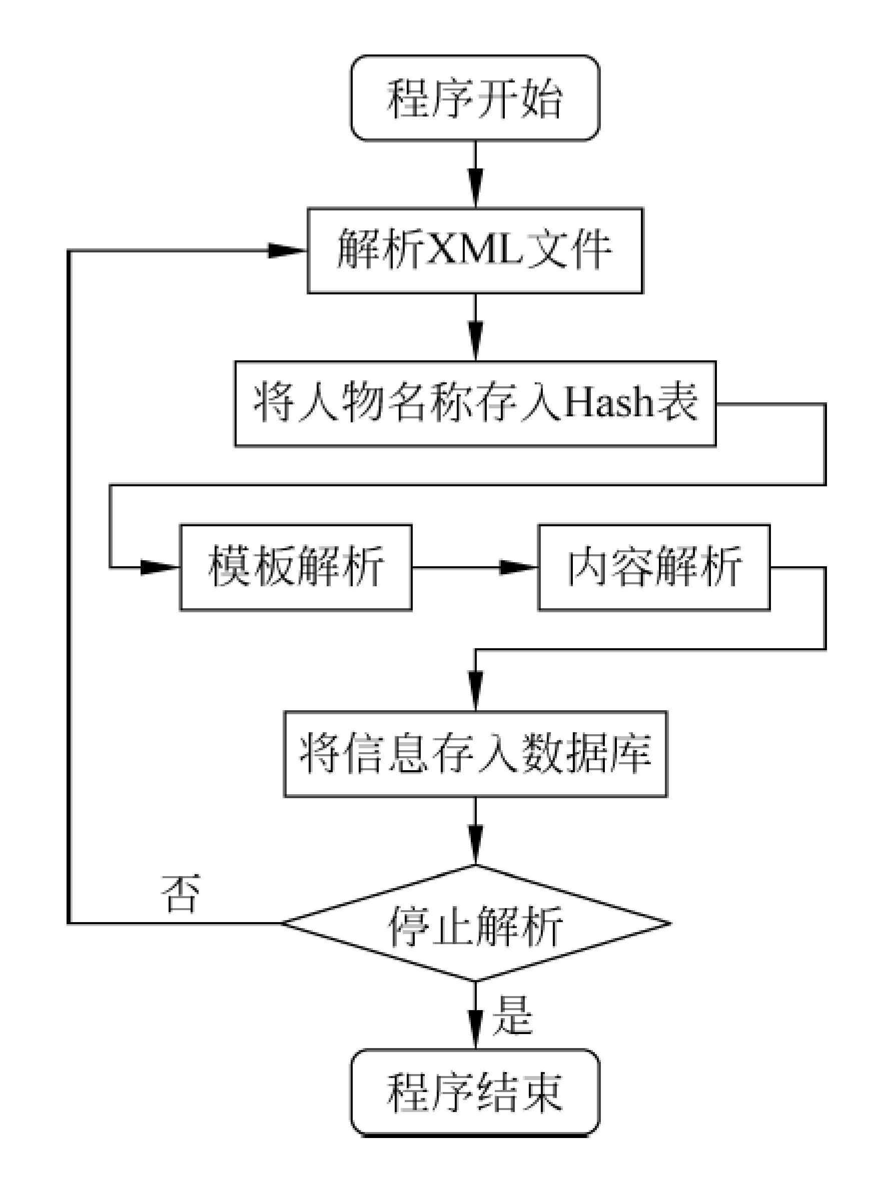


图 2.1　维基百科人物类数据获取示意图

在获取人物信息时，根据给定的人物名称列表，依次从 XML 格式数据包中找到对应的人物信息，并通过 SAX 解析出对应的人物信息。解析 XML 文档时，以人物名称为主键建立一个 Hash 表。由于维基百科通过重定向的方式避免了人物的重名，因此不必考虑人物名称重复的问题。通过模板解析器和正则表达式的方式从 Infobox 和正文内容中提取出人物页面 Infobox 信息，正文信息和锚文本信息，并把此信息作为人物实体信息存入对应的人物信息队列中。

2.5.2　新浪微博数据获取方法

新浪微博平台是一个开放的信息订阅、分享与交流平台。每条微博的字数最多不能超过 140 字，内容从兴趣爱好、饮食娱乐到政治时事均不受限制。不同于传统的社交媒体一对多的信息传播模式，微博平台的信息传播具有迅捷性和裂变性。鉴于微博的产生与传播特点，微博开放平台中包含海量的数据信息，如博主信息、微博信息、粉丝关系等。这些信息与关系有助于深入探索社会化媒体信息传播的机制与特点，非常具有研究意义。

新浪微博平台对使用者是开放的，每个使用者都可以使用新浪微博开放平台向外开发的一组 API 来获取指定格式的数据。API 是获取数据的接口，新浪微博 API 可供使用的接口有用户接口、微博接口、话题接口、好友分组接口、地理位置接

口及公共服务接口等。选取可以用来提供社会计算所需信息的接口，将它们分为以下三类。

(1) 基本信息资料接口，包括用户接口、用户标签接口等。

(2) 微博行为信息接口，包括微博接口、评论接口等。

(3) 用户关系信息接口，包括关系接口、好友分组接口等。

微博数据获取的流程图如图 2.2 所示。

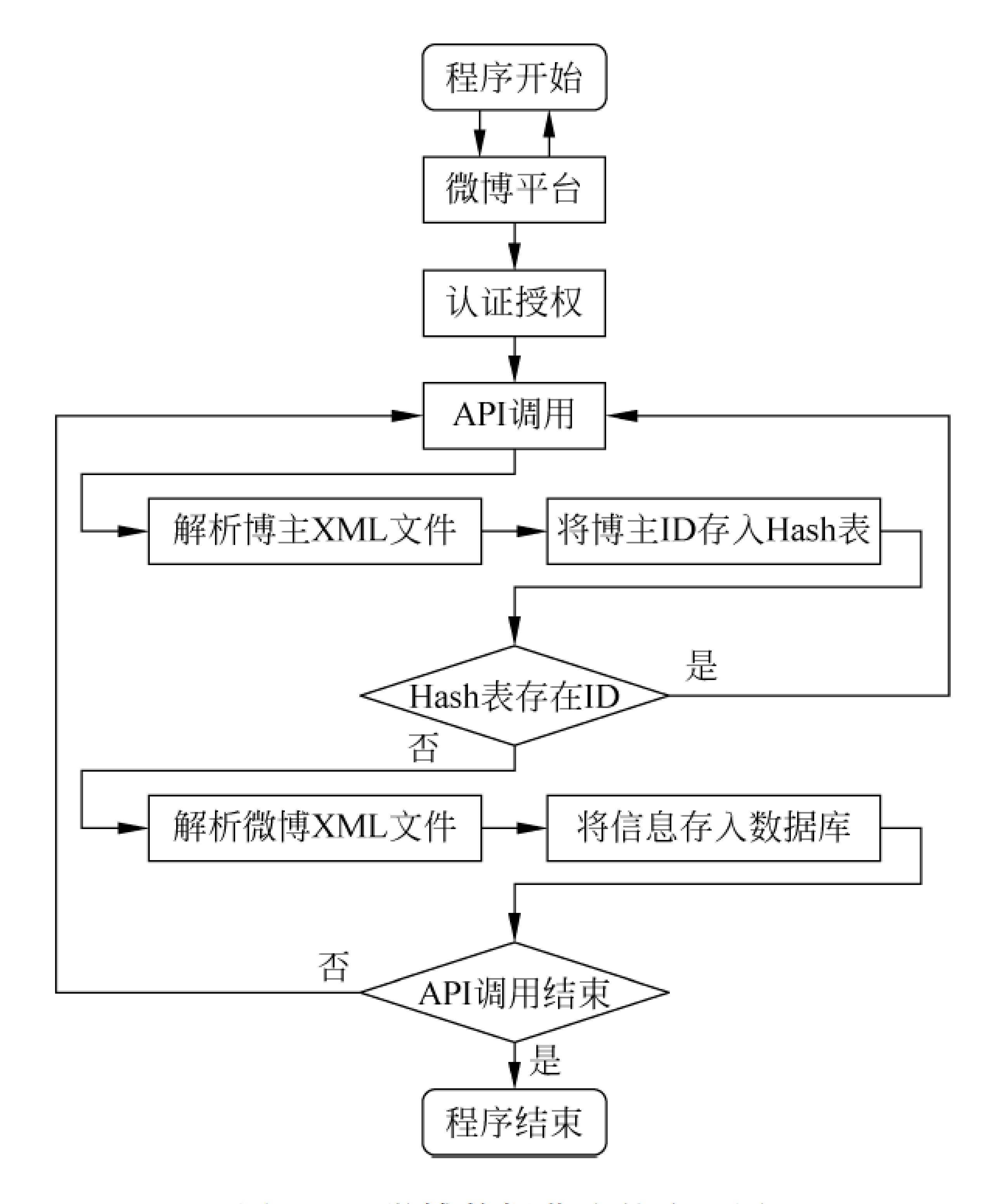


图 2.2 微博数据获取的流程图

微博数据获取的基本方式是通过 API 接口返回数据，默认以 XML 或 JSON 的格式返回博主信息。微博平台规定，当程序调用以上三类 API 接口时，需向服务器进行开放授权认证。

开放授权(OAuth)是一个开放标准，允许用户让第三方应用访问该用户在某一网站上存储的私密资源(如照片、视频、联系人列表)，而无须将用户名和密码提供给第三方应用。因此，OAuth 为新浪微博 API 提供了一个安全、高效的认证机制，其具体过程如下。

(1) 用户向新浪微博开放平台提出开发者服务申请，提交实名身份认证。

(2) 向新浪微博开放平台 OAuth 服务商提交创建应用请求，获得应用资料，并将其中的应用编号 App Key 和应用口令 App Secret 写入认证程序配置文件。

（3）利用新浪微博 SDK 提供的认证程序，向新浪微博服务器提交 API 使用申请，填写申请者微博账号、口令，获取第三方软件应用许可。

（4）申请成功后服务器在浏览器返回 URL 地址中提供一个由 32 位十六进制数组成的认证码 Access_code，用户将此认证码提交给认证服务器，服务器同意用户请求，向其颁发通过新浪微博授权的 API 调用令牌 Access_Token 与对应的密钥。

（5）用户利用此令牌作为参量调用相应的 API 接口。

通过上述 OAuth 认证登录新浪微博开放平台成功后，用户便可调用开放平台的各种接口，令牌使用期限为 24 小时，即超过试用期后需重新进行认证才能继续调用 API 接口。

授权认证完毕后，即服务器同意用户的接口调用请求，从 API 链接中打开一个输入流，从输入流中读取数据。其中，两个参数为博主 ID 和抓取的该博主的微博数，通过对两个参数的赋值，可以得到返回的 XML 格式的页面信息。

通过 SAX 将 XML 文件解析时，以每个博主 ID 为主键放入 Hash 表中，如果 Hash 表中存在该 ID 号，则停止解析。反之，继续解析每一条微博，包括博主账号、博主昵称、微博正文、转发数和评论数。解析完成后，将数据分别存放在数据库中，进一步判断是否继续调用 API。若继续调用，则循环解析程序，否则程序结束。

2.6　现有社会网络分析软件

社会网络分析的价值逐步显现出来，越来越多的人将注意力投入进来，社会网络分析软件随着人们的需求开始涌现。其中被广为使用的包括 UCINET、NetDraw、Pajek、NetMiner、StOCNET、Mage 等。

2.6.1　UCINET 软件

UCINET 基本上是最知名和最经常被使用的处理社会网络数据和其他相似数据的综合性分析程序。软件最初由美国加州大学欧文分校的一群网络分析者编写，现由斯蒂芬·波加提（Stephen Borgatti）、马丁·埃弗里特（Martin Everett）和林顿·弗里曼（Linton Freeman）组成的团队进行扩展和维护。UCINET 能够处理的原始数据为矩阵格式，它提供了大量数据管理和转化工具，但程序本身不包含网络可视化的图形程序，而是提供接口将数据和处理结果输出至 NetDraw、Pajek、Mage 和 KrackPlot 等软件进行作图。

UCINET 提供大量的网络分析指标的测量分析功能，包括凝聚子群分析、派系分析、中心性分析、个人角色分析和基于置换的统计分析等。另外，软件还包含

为数众多的基于过程的分析程序,如聚类分析、多维标度、二模标度(奇异值分解、因子分析和对应分析)、角色和地位分析(结构、角色和正则对等性)、拟合中心-边缘模型。此外,UCINET 提供从简单统计到拟合 P1 模型在内的多种统计程序。

UCINET 可以处理 32767 个网络结点,但从实际操作来看,当结点数在 5000～10000 之间时,一些程序的运行就会很慢。

2.6.2 NetDraw 软件

NetDraw 是由 Steve Borgatti 开发的开源工具软件,它通常用来对一模式和二模式网络进行可视化操作。NetDraw 可单独使用,也能被集成到 UCINET 中。它兼容多种文件格式,如 UCINET 的系统文件、DL 文本文件和 Pajek 的文本文件等;可以把网络的图形输出为 EMF、WMF、BMP 或 JPG 文件,也可以把数据输出到 Pajek 和 Mage 软件中。

2.6.3 Pajek 软件

Pajek 是一个特别为处理大数据集而设计的网络分析和可视化程序。Pajek 可以分析多于一百万个结点的超大型网络,并支持将大型网络分解成几个较小的网络,以便使用更有效的方法进一步处理。软件提供包括探测结构平衡和聚集性,分层分解和团块模型(结构、正则对等性)在内的基于过程的分析方法,但只包含少数基本的统计程序。Pajek 支持多种数据输入方式,包括 NET、CLU 和 VEC。网络文件(NET)中包含结点列表和弧/边(arcs/edges)列表,只需指定存在的联系即可,从而高效率地输入大型网络数据。软件使用的数据文件中可以包含指示行动者在某一观察时刻的网络位置的时间标志,因而可以生成一系列交叉网络,并对这些网络进行非统计性分析以及考查网络的演化。除了普通网络(有向、无向、混合网络)外,Pajek 还支持多关系网络、二模式网络(网络由两类异质结点构成),以及暂时性网络(网络随时间演化)。

2.6.4 NetMiner 软件

NetMiner 是一个把社会网络分析和可视化探索技术结合在一起的软件工具。使用者可以用可视化和交互的方式探查网络数据,最终找出网络的结构和潜在模式。NetMiner 采用的网络数据类型包括三种类型的变量:邻接矩阵、联系变量和行动者属性数据。NetMiner 具有与 Pajek 和 NetDraw 相似的高级图形特性,几乎所有的结果都是以文本和图形两种方式提交的。NetMiner 提供的网络描述方法和基于过程的分析方法也较为丰富,同时也支持包括描述性统计、ANOVA、相关和回归在内的一些标准统计过程。

2.6.5　StOCNET 软件

StOCNET 是个适用于社会网络高级统计分析的开放软件系统，它提供了一个应用多种统计方法的平台，每种统计方法以单独模块的形式嵌入其中。StOCNET 包含六个统计模块：

（1）BLOCKS，随机块模型。

（2）ULTRAS，使用超度量估计潜在的传递性结构。

（3）P2，拟合指数随机图 P2 模型。

（4）SIENA，纵向网络数据的分析。

（5）ZO，确定随机图统计量的分布概率。

（6）PACNET，构造和拟合基于偏代数结构的结构模型。

2.7　本章小结

社会化媒体不仅给广大的用户提供了分享和交流的新型媒体平台，而且也为社会计算的研究者们提供了丰富的媒体大数据工厂。如何快速成为新型媒体的主力军，分析社会媒体的多样性呈现方式以及平台所传递的信息对社会媒体的发展和研究都至关重要。首先，本章从社会媒体的定义入手，阐述了社会媒体所表现出的独特特征，揭示了社会媒体在互联网中的重要性。其次，对多态呈现的社会媒体进行了分类，分析了每种分类的呈现特点和优势，并列举了有代表性的社会媒体网站。第三，本章从社会媒体研究者的角度出发，详细介绍了两种典型的媒体大数据的获取方法。最后，本章还介绍了目前现有的社会网络分析软件，分别从各自的适用范围、功能和特点等方面呈现了对社会计算的研究辅助作用。

思考题

1. 什么是社会化媒体？它包括哪些特征？根据你的体验，谈谈你对社会媒体的认识。

2. 根据社会化媒体的呈现方式，试描述社会化媒体都有哪些类别？从这些类别中，试选出两个以上你使用的社会化媒体，谈谈你使用的感受。

3. 查阅相关资料，浅析主流的社会化媒体的发展现状以及所面临的问题。

4. 根据社会化媒体的数据获取方法，试着设计一个网络媒体数据的爬取程序，并分析所爬取数据的社会化特征。

5. 利用一些经典的社会网络的数据集，试选择 1～2 个社会网络分析软件，对数据集进行网络分析。

第 3 章

基于网络结构的社区发现

本章学习目标

- 理解社区发现的目的和概念
- 掌握经典的非重叠社区发现算法
- 理解重叠社区发现的意义和算法思想

3.1 非重叠社区发现

社区发现具有重要的理论意义和应用价值，它吸引了包括计算机、生物、社会学、物理学、数学等诸多不同领域的研究者进行研究。从 2002 年开始，研究者针对不同的问题和领域提出不同的解决思路，研究成果已在不同领域的权威国际期刊和重要学术会议论文集上发表。

3.1.1 传统算法

1. 图分割法

将网络看作一个图，社区看作是密集的子图结构，就可以使用计算机领域经典的图分割法来解决社区发现问题。

图分割法的目标是把图中的结点分为 n 个预定大小的群组，并使这些群组之间的边数最小。通常利用迭代对网络进行划分：先将网络最优划分为两个子网络，再重复对子网络进行最优二分，直到最终得到 n 个子网络(即 n 个社区)。

使用较多的两个算法是 Kernighan-Lin 算法和谱平分法。两种算法的主要局限在于不能保证迭代二分就能得到正确的划分，而且缺乏有效的二分停止条件。另外，Kernighan-Lin 算法还需要预先知道两个子网络的大小。

2. 层次聚类

通常有许多网络是具有层次结构的，几个小社区被包含在较大的社区里，这些社区又被包含在更大的社区中。在这样具有层次结构的社区中，使用层次聚类的方法获取社区结构可获得较好结果。

所有的层次聚类方法都要首先定义一个计算结点相似度的方法，然后根据此方法计算任意两结点间的相似度，形成一个相似度矩阵。而后根据计算方向的不同，层次聚类可具体分成凝聚法和分裂法。

凝聚法：将每个结点视为一个初始社区，根据相似度从强到弱逐步重新连接各结点，形成树状图(Dendrogram)，如图 3.1 所示，根据需求对树状图进行横切，获得社区结构。主要步骤如下。

(1) 移除网络中的所有边，得到有 n 个孤立结点的初始状态。

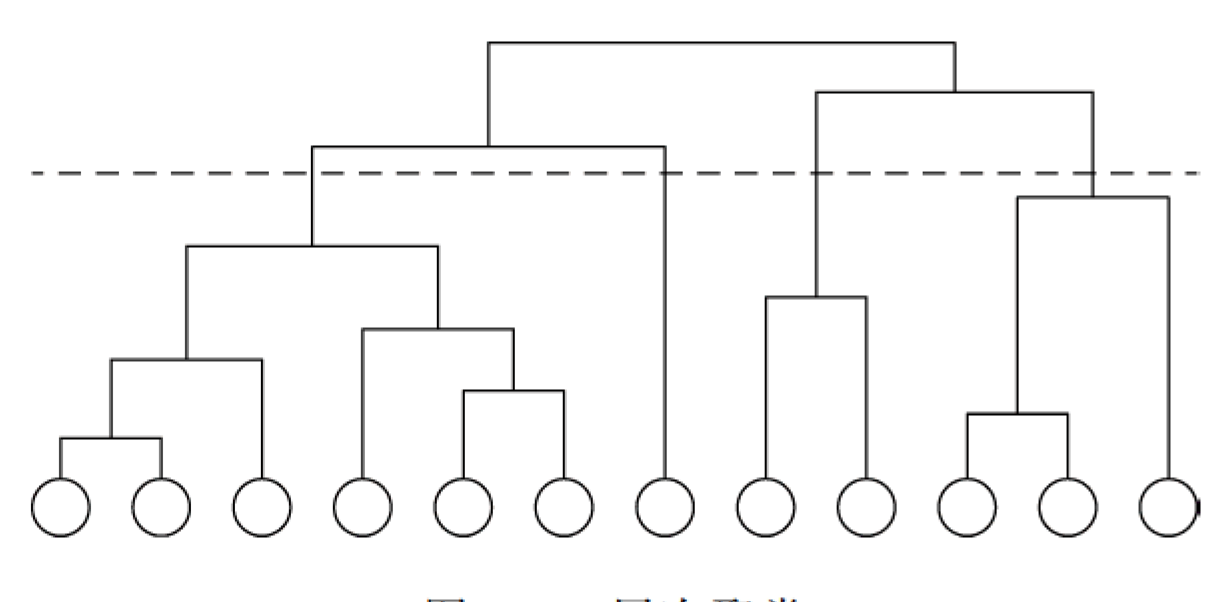


图 3.1　层次聚类

(2) 计算网络中每对结点的相似度(不考虑两结点是否相连)。

(3) 根据相似度从强到弱连接相应结点对，形成树状图。

(4) 根据实际需求横切树状图，获得社区结构。

分裂法：找出相互关联最弱的结点，删除它们之间的边，通过这样的反复操作将网络划分为越来越小的组件，最终依然连通的网络构成社区。

层次聚类法的优点在于不需要指定网络的社区个数或社区规模。但该方法并不能确定网络的最优划分，而且非常依赖于结点相似度的衡量标准。其聚类结果有可能会将某些重要结点划分为单独的社区，从而不能正确划分网络的外围结点。

3. 其他聚类

通过给每个网络结点分配一个合理的 K 维坐标，可以把社区发现问题转换为传统的空间点聚类问题，然后就可以采用K-means 等经典聚类算法将这些新生成的空间点聚类。早在 1970 年，Hall 针对图分割问题提出了加权二次型变换算法。该算法能够将网络投影到一维空间，使得网络中连接紧密的结点在一维空间中的位置相对较近，而连接稀疏的结点在一维空间中的位置相对较远。类似地，

Donetti 和 Munoz 在 2004 年提出了一种结合谱方法和空间点聚类方法的复杂网络聚类算法。算法通过计算拉普拉斯矩阵的 K 个最小特征向量将网络映射到 K 维空间中，然后采用某种基于距离的空间点聚类算法聚类网络结点。

3.1.2 分裂算法

分裂算法通过识别并切断连接不同社区结点的边来发现社区结构，算法的关键在于找到连接不同社区的边的属性特征以便在图中识别它们。实际上，分裂算法与层次聚类的方法比较类似，均是以切断一定条件的边的方式进行，区别在于，分裂算法移除的是社区之间的关联边，而这些边上两点的相似度不一定很低。其中最著名的算法就是 Girvan-Newman 算法，根据以下假设：社区之间所存在的少数几个连接应该是社区间通信的瓶颈，是社区间通信时通信流量的必经之路。如果考虑网络中某种形式的通信并且寻找到具有最高通信流量（例如最小路径条数）的边，该边就应该是连接不同社区的通道。Girvan-Newman 算法这样迭代删除边介数（Edge Betweenness）最大的边。

除此之外，Tyler 等人在 2003 年将统计方法引入基本的 GN 算法，提出一种近似 GN 的算法。算法采用蒙特卡洛方法估算出部分连接的近似边介数用以取代精确边介数，牺牲聚类精度来提高计算速度。2004 年，Radicchi 等人提出以连接聚类系数取代 GN 算法的边介数。算法基于社区间连接应该很少出现在短回路（如三角形或四边形）中的假设，把连接聚类系数定义为包含该连接的短回路数目。而社区间连接的连接聚类系数应小于社区内连接的连接聚类系数，因此算法在迭代过程中不断删除具有最小连接聚类系数的边。该算法的最大局限性是：不适合处理短回路很少甚至没有的复杂网络。

3.1.3 基于模块度的方法

为了衡量社区发现的结果，Newman 和 Girvan 在 2004 年提出模块度评价函数（Q 函数）。Q 函数的定义为社区内实际连接数目与随机连接情况下社区内期望连接数目之差。Q 的计算公式如式（3-1）所示。

$$Q = \sum_{S=1}^{K}\left[\frac{m_s}{m} - \left(\frac{d_s}{2m}\right)^2\right] \tag{3-1}$$

其中 K 表示社区个数，m 表示网络连接总数，m_s 表示社区内连接总数，d_s 表示社区 s 中结点度之和。Q 位于 0 到 1 之间，取值越大代表结果越好。

Newman 同年提出了第一个基于模块度优化的算法，算法初始化时将每个结点看作是一个社区，在不断的迭代中选择使 ΔQ 最大化（不一定为正）的目标合并两个社区，直到网络中只剩一个社区。最终在形成的层次聚类树中选取令 Q 最大的社区划分作为结果。

Guimera 和 Amaral 在 2005 年提出了基于模拟退火的模块度优化算法(Genetic Algorithm,GA),并应用到新陈代谢网络分析中。GA 算法通过将结点移动到其他社区、交换不同社区内的结点、分解或合并社区三种策略产生新的候选解,利用候选解的 Q 值来进行评价,并采用模拟退火策略的 Metropolis 准则决定是否接受该候选解。算法的准确度极高,但效率较低,不适用于大型网络。

针对模块度优化的算法非常多,许多传统的优化算法都可以被用来对 Q 值进行优化,相关的研究成果也很多。但值得注意的是,对于大规模的复杂网络,基于优化 Q 函数的复杂网络聚类算法倾向于找到粗糙而不是精细的社区结构,意味着这些未必能够找到网络中真实存在的全部社区结构。

3.1.4　动力学算法

分析网络的动力学过程,可以发现网络的结构属性,进行社区发现。基于动力学的社区发现算法主要是随机游走方法和同步方法。

随机游走方法是基于以下思想:如果存在很强的社区结构,那么随机游走器(Random Walker)会在社区内部停留更长的时间,因为社区内部的边密度比较高。

例如,有的学者提出了基于马尔可夫随机游走模型的启发式符号网络聚类算法(Finding and Extracting Communites,FEC)。FEC 算法所采用的基本假设是:从任意给定的社区出发,网络中的随机游走过程达到起始社区内结点的期望概率将大于达到起始社区外结点的期望概率。基于该启发规则,FEC 算法首先计算出在给定时刻随机游走过程到达所有结点的期望转移概率分布,进而根据该分布的局部一致性——同社区结点具有近似相同的期望转移概率分布——识别出各个不同的网络簇。与现有方法相比,FEC 算法在时间和识别精度方面表现出了更好的性能,尤其适合处理噪声高和社区结构不明显的复杂网络。该算法的参数是随机游走的步长,步长的设置会影响最终的聚类结果。通过实验分析,FEC 算法给出了步长设置的经验值,建议取值区间为。其中 6 表示复杂网络中两点间的平均距离(大多数网络都满足六度分离理论),20 表示网络的直径(WWW 是迄今最大的复杂网络,研究表明其直径为 19)。但是 FEC 算法没有从理论上给出一种针对不同网络设置最优参数的方法。

3.1.5　局部社区发现算法

在许多超大型且动态变化的网络结构中,全局的社区发现算法往往是不可行的,针对这个问题,研究人员提出一些寻找网络中局部社区结构的算法。比较有代表性的包括 Hub 算法和 BB 算法。

Hub 算法是 Costa 等在 2004 年提出的,其中心思想为:在许多实际网络中,

社区是以一些具有最大度的 Hub 结点为核心产生的，这些核心结点会不断地吸引周围的结点，因此会以它们为中心形成各个社区。Hub 算法最大的局限在于必须知道社区的数目，且要求这些社区的直径是相等的，否则就很容易出错。

BB 算法继承了 Hub 算法的部分思想，从已知结点出发，通过扩展传播寻找结点所在的社区结构，并通过对定义的暴露度的增长限制来控制扩散，最终得到结点所在社区结构。

3.1.6 几种经典社区算法

1. Kernighan-Lin 算法

Kernighan-Lin 算法为网络的划分引入一个增益函数，并利用贪婪搜索得到增益函数最大的网络划分。增益函数的定义为两个子网络内部边的数量减去子网络间边的数量，将待分割的网络随机划分成指定大小的两个子网络，不断地交换两个子网络的结点来对增益函数进行优化。

首先，将网络中的结点随机地划分为已知大小的两个子网络。在此基础上，考虑所有可能的结点对，其中每个结点对的结点分别来自两个子网络。对每个结点对，计算如果交换这两个结点可能得到的 Q 的增益 $\Delta Q = Q_{交换后} - Q_{交换前}$，然后交换最大的 ΔQ 对应的结点对，同时记录交换以后的 Q 值。规定每个结点只能交换一次。重复这个交换过程，直到某个社团内所有的结点都被交换一次为止。需要注意的是，在结点对交换的过程中，Q 值并不一定是单调增加的。不过，即使某一步的交换会使 Q 值有所下降，仍然可能在其后的步骤中出现一个更大的 Q 值。当交换完毕后，便找到上述交换过程中所记录的最大的 Q 值。这时对应的就是最终结果。

在整个搜索过程中，KL 算法只接受更好的候选解，而拒绝所有较差的候选解，因此它找到的解往往是局部最优而不是全局最优解。KL 算法最大的局限性在于它需要先验知识(社区的个数或社区的平均规模)来产生一个较好的初始结构，因为该算法对初始解非常敏感，不好的初始解往往导致缓慢的收敛速度和较差的最终解。

2. 谱平分法

一个有 n 个结点的无向图的 Laplace 矩阵是一个 $n \times n$ 维的对称矩阵 L。其中，L 的对角线上的元素 L_{ii} 是结点 i 的度，其他非对角线上的元素 L_{ij} 则表示结点 i 和结点 j 的连接关系。如果这两个结点之间有边连接，则 L_{ij} 值为 -1，否则为 0。也可以将矩阵 L 表示成 $L = K - A$，其中，K 是一个对角矩阵，其对角线上的元素就对应各个结点的度，A 则为该网络的连接矩阵。L 矩阵所有的行与列的和都为 0，因此，该矩阵总有一个特征值为 0，且其对应的特征向量为 $l = (1, 1, \cdots, 1)$。

从理论上可以证明，不为零的特征值所对应的特征向量的各元素中，同一个社团内的结点对应的元素是近似相等的。这就是谱平分法的理论基础。

当一个网络中仅存在两个社区，此时该网络的 Laplace 矩阵 L 仅对应两个对角矩阵块。对一个实对称的矩阵而言，其非退化的特征值对应的特征向量总是正交的。因此，除最小特征值 0 以外，矩阵 L 其他特征值对应的特征向量总是包含正、负两种元素。当网络由两个社区构成时，就可以根据非零特征值相应的特征向量中的元素来对应网络的结点进行分类：所有正元素对应的那些结点都属于同一个社团，而所有的负元素对应的结点属于另一个社团。由此可以根据网络的 Laplace 矩阵的第二小的特征值 K_2 将其分为两个社团。这就是谱平分法的基本思想。

当网络的确是近似地分成两个社团时，用谱平分法可以得到非常好的效果。但是，当网络不满足这个条件时则不行。而实际上，第二小特征值 K_2 可以作为衡量谱平分法效果的标准：它的值越小，平分的效果就越好。

一般情况下，计算一个 $n\times n$ 矩阵的全部特征向量的时间复杂度为 $O(n^3)$。但是在大多数情况下，实际网络的 Laplace 矩阵是一个稀疏矩阵，可以用 Lanczos 方法快速计算主要的特征向量。该方法的时间复杂度大致为 $O(m)$，其中，m 表示网络中边的条数。这样，计算的速度可以得到明显的提高。但是，如果不能很快将 K_2 从其他特征值中分离出来，算法就可能在一定程度上有所减慢。换句话说，当网络很明显地分成两个社团时，该算法的速度非常快，否则该算法就未必很有效。

3. GN 算法

GN 算法是 Girvan 和 Newman 于 2002 年在 PNAS 上发表论文提出的，该论文不仅为网络社区结构的研究拉开了序幕，同时也提出了一种基于边介数(Edge-Betweenness)的分裂式层次社区发现算法，边介数是指网络中的某边是网络中任意两点的最短路径中边的个数，边介数的概念是从边在社区中所起的作用和位置出发的，若某边的边介数很大时，说明这条边充当了多个点之间的桥接，那它是两个社区之间的边的可能性也最大，因此，移走最大边介数的边将可以分离出两个社区。也就是说，若两个社团经过一条边相连，则这两个社区结点间的最短路径通过此边的次数最多，即该边边介数最大，通过删除该边，两个社区即可分开。

GN 算法就是利用了这个原理，它首先将整个网络中所有结点看作是一个社区，然后移除边介数最大的边，社区被分裂，接着在分裂的各自社区中继续移除边介数最大的边，直到无法再移除边或是每个结点自形成一个社区时停止，清楚地看到这个迭代的过程形成的是一棵分裂树，当选择分裂树的不同地方分割时，GN 算法就会形成相应的社区结构。

具体算法如下：

（1）计算网络中所有边的介数。

（2）移除介数最高的边。

（3）重新计算所有受影响的边的介数。

（4）重复步骤(2)，直到每个结点就是一个退化社团为止。

分裂过程中算法可在任意时刻终止，并得到当前结果作为发现的社区结构。因此循环终止的条件可以有很多种，例如限定划分的社区结构、得到的社区结构性质（强连通、弱连通）、模块度要求等。

使用 GN 算法可以较好地发现网络存在的社区结构，算法对存在孤立结点的网络、全连接社区、无权图、高内聚网络等特殊形式，均表现出良好的鲁棒性。

GN 算法一经提出，就受到广泛关注，在社区发现算法中占有相当重要的位置。

4. W-H 算法

Wu 和 Huberman 提出一种基于电阻网络电压谱的快速谱分割法，算法复杂度只有 $O(m+n)$，是一种线性算法复杂度的算法。利用该算法不仅可以求出网络的社区结构，还可以在不考虑整个网络社区结构的情况下，寻找一个已知结点所在的整个社区，而无须计算出所有的社区，这是其他很多算法都无法实现的。但 W-H 算法的缺点在于，在没有预先知道社区数目的情况下无法使用。

假设图 $G=(V,E)$ 可以分为两个社团 G_1 和 G_2，且已知结点 A 和 B 分别属于这两个社团。令结点 A 为源结点，电压值为 1，而结点 B 为终结点，电压值为 0。此时，网络中的每条边都视为一个阻值为 1 的电阻。整个网络就可以看成一个电阻网络，从而可以利用 Kirchhoff 定理求各个几点的电压值。然后，选取一个电压阈值 $V(0<V<1)$。若结点 i 的电压值 $V_i>V$，则认为它属于源结点 A 所在的社团，反之则属于终结点 B 所在的社团。实际上，可以利用谱线图来记录电压值：在 0～1 的范围内，将电压值从小到大进行排列，然后用不同位置的谱线图来记录电压值。这样构成的谱线图就称为电压谱。然后选取某个阈值，认为该阈值左边的谱线相应的结点属于一个社团，而右边的那些结点就属于另一个社团。

算法采用统计法来选取分别位于两个不同社区内的结点。社团内部的联系比较紧密，而社团之间的结点的联系就相对比较松散。因此，只要两个点的距离大于 2 就极有可能位于不同的两个社团。在这个思想下，算法可以分两步进行：首先，随机选取若干对距离大于 2 的结点，例如 50 对，并对每一个结点对利用电压谱将网络划分为两个社团，由此就可以得到 100 个社团。然后，从网络中任意选择一个结点作为参考结点，看其他结点与该结点同时出现在一个社团内的概率，如果超过 50%，则认为这两个结点属于同一个社团，反之则认为他们属于两个不同的社团。

5. FastNewMan 算法

在动辄包含几百万个以上结点的大型网络中，传统的 GN 算法就不能满足要求。基于这个原因，Newman 在 GN 算法的基础上提出了一种快速算法，它可以用于分析结点数达 100 万的复杂网络。这种快速算法实际上是基于贪婪算法思想的一种凝聚算法。算法如下：

（1）初始化网络，将每个结点看作是一个独立社团。初始的 e_{ij} 和 a_i 满足 $e_{ij}=1/2m$，如果结点 i 和 j 之间有边相连，其他 $a_i=k_i/2\ m$ 其中 k_i 为结点 i 的度，m 为网络中总的边数。

（2）依次合并有边相连的社团对，并计算合并后的 Q 值增量：$\$ Q=e_{ij}+e_{ji}-2a_ia_j=2(e_{ij}-a_ia_j)$。根据贪婪算法的原理，每次合并应该沿着使 Q 增大最多或者减少最小的方向进行。该步的算法复杂度为 $O(m)$。每次合并以后，对相应的元素 e_{ij} 更新，并将与 i、j 社团相关的行和列相加。该步的时间复杂度为 $O(n)$。因此，步骤(2)总的时间复杂度为 $O(m+n)$。

（3）重复执行步骤(2)，不断合并社团，直到整个网络都合并成为一个社团。最多要执行 $n-1$ 次合并。

该算法总的算法复杂度为 $O((m+n)n)$，对于稀疏网络则为 $O(n^2)$。整个算法完成后可以得到一个社团结构分解的树状图。再通过选择在不同位置断开可以得到不同的网络社团结构。在这些社团结构中，选择一个对应着局部最大 Q 值的，就得到最好的网络社团结构。

3.2　重叠社区发现

作为复杂理论的一个重要支撑，研究网络已经被证明是理解许多自然和人工系统的结构与功能的最有效的内容，而且，复杂网络的最普通的特征就是社区结构，因此，发现社区结构并分析是了解现实生活中各种网络组织结构的一种很重要的方法，在生物学、计算机科学以及社会学等领域都有着广泛的应用。

3.2.1　重叠社区发现

通常，社区指的是一组结点在组内比网络其他部分的结点连接得更紧密，模块和社区反映了网络元素之间的拓扑关系和代表功能实体。比如说，社区可能是社会网络中一群相关个体的群体，或是一组处理同一个主题的网页集合，也可能是在循环代谢网络中的一条生物化学链。因此，复杂网络中的社区确认是非常重要的，然而它也有许多难题需要研究者们去钻研。

摆在研究者面前的难题中，社区的重叠性是其中重要的一个问题。社区的重

叠性是指网络中的结点经常属于不止一个模块或社区，也就是属于多个社区，这就形成了重叠社区。事实上，社区的重叠性是个显而易见的网络特性，比如说，在社会网络中的每个个体，网络中每个个体都可能依附于多种关系，比如说：家庭、朋友、专业、兴趣爱好等，假如按照这些依附关系去进行个体划分，就会清楚地发现每个个体都可以属于不同的社区。然而，重叠性的存在给社区发现带来了困扰，重叠性会降低已发现社区的质量，更进一步来说，重叠性会隐藏一些重要的信息，由此会经常产生结点的误判。

图 3.2 是描述的重叠社区的一个例子，从图 3.2 中可以看到，浅色结点分别组成了三个独立的社区，深色结点在多个社区之间担任着重要的角色，它们属于多个社区，并和这些社区保持着同等重要的关系，因此这些深色结点被视作重叠结点。

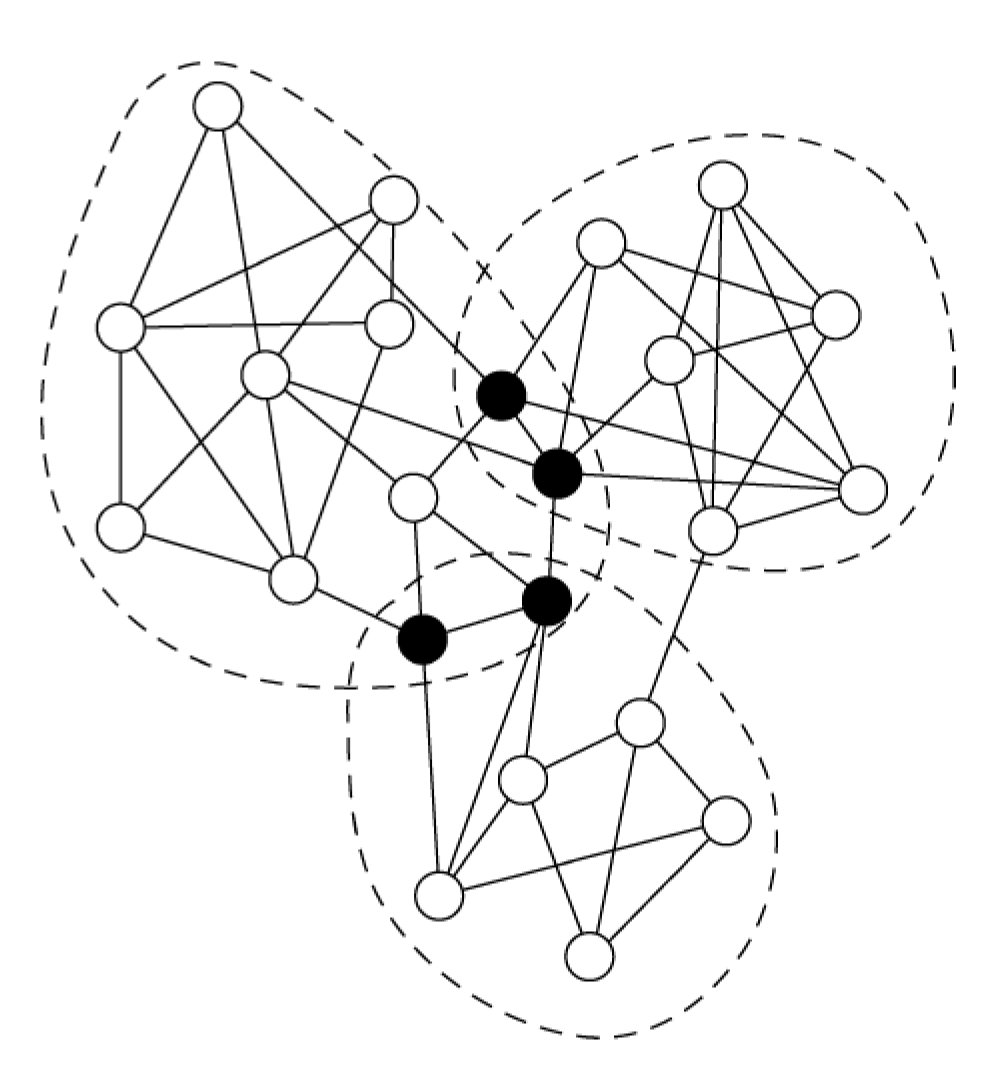


图 3.2 重叠社区例子

（资料来源：Lancichinetti，Fortunato. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New J Physics，2009，11(3)：033015.）

3.2.2 重叠社区发现算法分类

1. EAGLE 算法——基于重叠社区模块度优化算法

Newman 和 Girvan 于 2004 年提出了一个衡量网络社区结构优劣的量化标准——模块度(Modularity)函数，模块度函数是通过考查结点的度分布来测量现实的社会网络中社区划分粒度的方法，其根本思想是将划分后的社区结构中的结点之间的连接情况与相应的零(或随机)模型的连接期望进行比较以确定划分的质量。零(或随机)模型是指与现实的网络具有相同的性质(如相同边数或度序列)，

而在其他方面完全随机的随机图模型。自模块度函数被提出后，其成为当前社区发现算法中应用最为广泛的判定社团关系强弱的指标，Newman 在模块度提出的同一年也提出了基于模块度优化的社区发现 FastNewman 算法，该算法的核心思想是取模块度最大的社区划分，其时间复杂度是 $O(mn)$。尽管基于模块度优化的社区发现方法已成为复杂网络社区发现领域中的主流方法之一，例如：模拟退火算法、数学规划方法等，然而研究者们已证明优化模块度的方法是个 NP 难问题，但是，这些近似算法在某种意义下可以得到复杂网络的非重叠的社区划分。因此，对于具有重叠性的社区，一些研究者类推模块度的定义给出具有重叠性的模块度定义，更进一步延伸模块度优化的方法到重叠社区发现。

有的学者人直接从 Newman 的非重叠社区发现算法的模块度定义出发，给出一种简单的针对重叠社区的模块度定义，这个定义的前提条件是网络的社区结构已经得到划分，并且允许网络中的结点可以同时属于多个社区，定义如公式(3-2)所示：

$$EQ = \frac{1}{2m}\sum_{i}\sum_{v\in C_i, w\in C_i}\frac{1}{O_v O_w}\left(A_{vw} - \frac{k_v k_w}{2m}\right) \tag{3-2}$$

其中，公式(3-2)中 O_v 表示结点 v 所属于的社区的个数，m 是实际网络的总边数，A_{vw} 对应结点 v 和 w 的链接关系，在无权值的网络中，A_{vw} 用 1 和 0 代表链接和不链接关系，在有权值的网络中，A_{vw} 可用边 e_{vw} 权值代表链接关系，k_v 表示结点 v 的度数。从公式中不难看出，当每个结点只能属于一个社区时，这个公式得出的 EQ 就退化为非重叠社区的模块度 Q 值。同样，该定义具有和模块度 Q 值类似的性质，即当所有结点属于一个社区时，模块度 Q 的值为 0；EQ 值的大小反映出网络重叠结构的明显度，即当 EQ 模块度值越大时，其所表示的网络重叠社区结构越明显。

自此之后，Shen 在文中提出了一种同时发现满足层次性和重叠性的分层凝聚的 EAGLE 算法。该算法与传统的基于结点之间不断聚合的社区发现方法不同，其处理的对象是网络中的极大团，通过极大团的不断聚合来形成网络社区的划分。所谓的极大团(Maximal Clique)是指极大团所构成的结点集合不是任意其他结点集合的子集，也就是这个结点集合是包含这些结点的最大的团体，且不能再分割为更小的团体。EAGLE 算法分为两个阶段，第一个阶段是通过搜索网络中极大团的方法生成网络的树状图，第二个阶段是选择合适的位置断开生成树，断开生成树的办法就是通过上述测量公式的重叠模块度值，EQ 值越大意味着分割越好，由此得到相应的社区划分结构。在第一个阶段中，首先 EAGLE 算法采用成熟的 Bron-Kerbosch 算法找出网络中所有的极大团。为了避免次大团对于算法搜索过程所产生的误导作用，算法会通过设置相应的阈值 k 去屏蔽掉一些小规模(小于 k)的极大团，阈值 k 的大小决定了被忽略的极大团的多少，k 值越大意味着屏蔽的极大团

数目越多，反之，一些次大团将会被保留下。其次，在确定好极大团之后，EAGLE 算法通过测量任意两个极大团之间的相似性，并选择相似性最大的极大团进行合并形成新的大团，不断重复这个步骤，直到只有一个社区为止。测量社区之间的相似性的办法就是采用一定的变形，针对两个社区的重叠模块度的定义。

假设网络中有 n 个结点，EAGLE 算法在第一个阶段的时间复杂度为 $O(n^2+(h+s)s)$，在第二阶段的时间复杂度是 $O(n^2s)$，其中，s 是指在第一个阶段中所搜索出的极大团的个数，h 是指在第一个阶段中相邻的极大团的成对个数。因此，综合两个阶段来看，EAGLE 算法的时间复杂度是 $O(n^2s)$。

2. CPM 算法——基于派系过滤算法

有的学者提出派系过滤算法，派系过滤算法的本质特征是重新给出社区的新定义或者说是关于社区的另一个前提条件，其认为典型的社区应是一些全连通的完全子图，这些完全子图通常被称之为团，也称作派系(Clique)，这些团(或派系)则表现出团内部的边连接密度较高，而在团之间的边形成团的可能性较小的特性，因此，派系过滤算法的主要目的就是找到这些紧密相连的完全团。通常，由 k 个结点组成的完全子图叫做 k-派系(k-Clique)，如果两个 k-派系有 $k-1$ 个共享结点，则称它们是相邻的派系。更进一步说，派系过滤算法的目的是找出网络中最大的全连通子图(或称为派系)，这些全连通子图之间共享的结点就是重叠结点，具有重叠结点的派系过滤算法的社团示意图如图 3.3 所示。

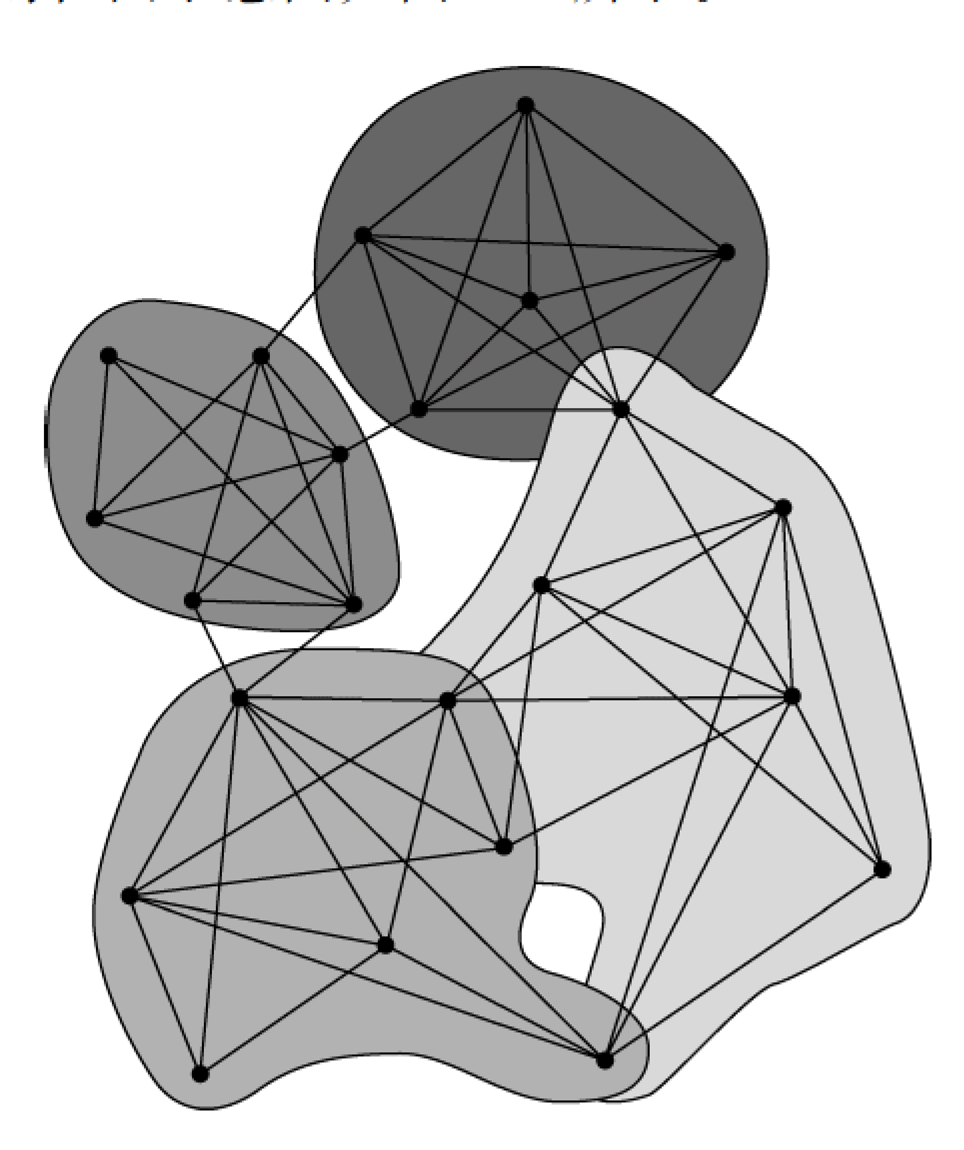


图 3.3　有重叠结点的派系示意图

(资料来源：Palla，Derenyi，Farkas. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature，2005，435：814-818.)

作为派系过滤算法的起源算法，CPM(Clique Percolation Method)算法的主要思想是首先从网络中找出所有大小为 k 的团，然后再把找出的每个 k 团作为结点构建一个新图，假如两个 k 团共享 $k-1$ 个结点时，那么新图中两个对应的结点之间才会有边，最后新图中每个连通子图所对应的 k 团集合才构成了一个社区。因此，可以看到，CPM 算法是通过合并全连通子图的方法来构建社区的，同时，由于一个结点可能会同时属于多个 k 团，所以 CPM 确定的社区自然会出现重叠，这意味着 CPM 可实现重叠社区的发现。

CPM 算法的实施过程简要描述：首先，找出网络中所有的完全子图，且保证这些完全子图不是更大的完全子图的子图，这些完全子图就是派系过滤算法中所说的派系(Clique)。事实上，派系与 k-派系的本质区别是派系可以是更大完全子图的子集。一旦派系确定，接下来就可形成派系与派系之间的(Clique-Clique)重叠矩阵。重叠的对称矩阵中每一行(或列)表示一个派系，矩阵元素等于两个对应的派系的共同结点，对角线上的元素等于派系的大小。对于给定 k 值的 k-派系社区等价于相互连通的相邻派系连接至少具有 $k-1$ 个共同结点，而这些部分是可以在对称矩阵中得到，即将对称矩阵中非对角线上值小于 $k-1$ 的元素以及对角线上值小于 k 的元素用 0 代替，用 1 代替剩下的元素，由此就完成了派系之间的重叠矩阵组成的分析转化，转化后的矩阵中剩下的为 1 的部分就是 k-派系社区，该分析过程如图 3.4 所示。

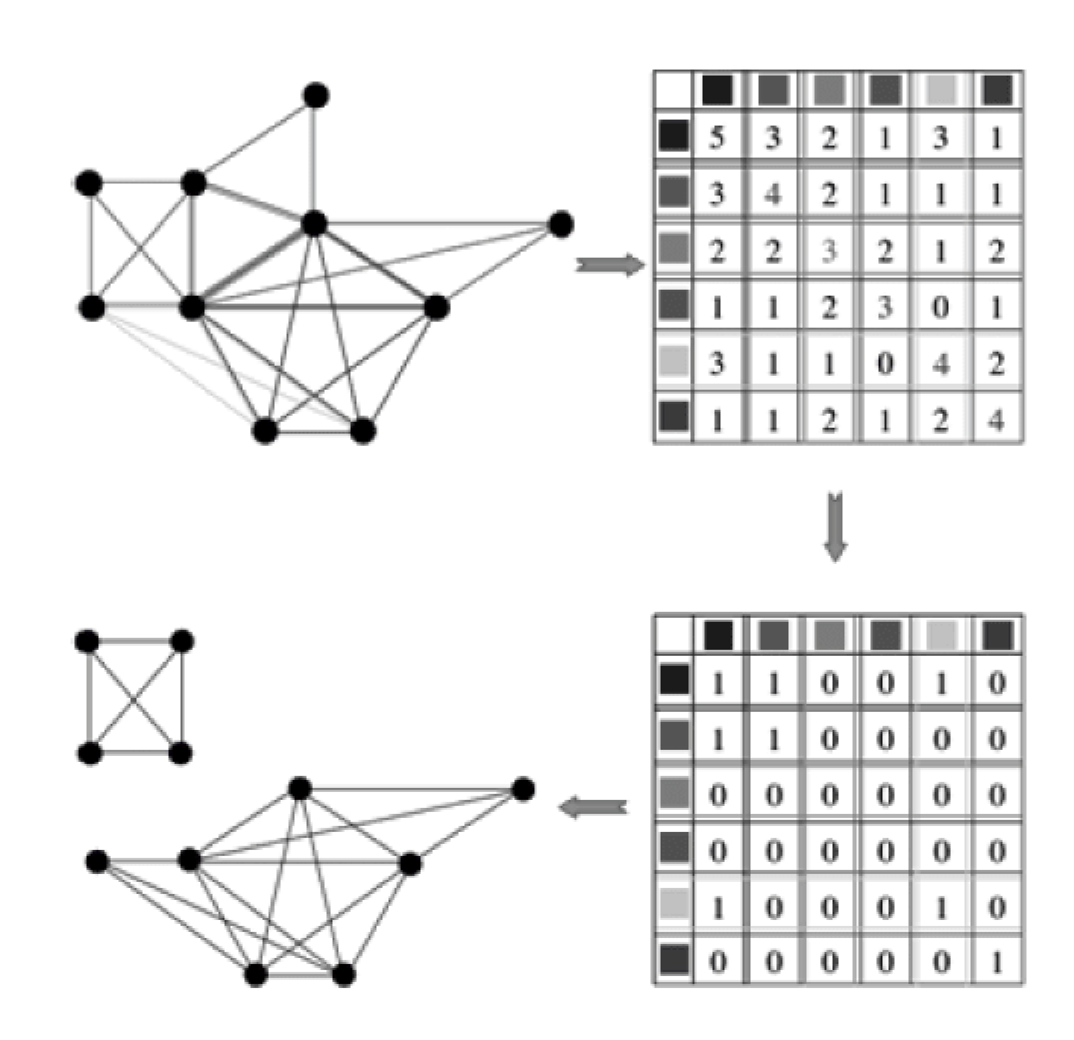


图 3.4　k-派系社区的分析过程

(资料来源：Palla，Derenyi，Farkas. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature，2005，435：814-818.)

尽管 CPM 算法对重叠社区的发现一般来说是非常有效的，然而，由于 CPM 算法所基于的新的社区定义或者说是前提假设，因此，其不可避免地有以下缺点：

（1）计算网络中的全部 k-团非常耗时，其时间复杂性近似为指数阶。

（2）当网络中的全连通子图非常少时，它就难以体现优势了。

（3）参数 k 值确定困难，不同的 k 值将会得到不同的网络社团结构。

3. CONGA 算法——基于分裂介数的 GN 算法

2007 年在第 11 界欧洲国际数据挖掘原理与发现会议（PKDD）上，Gregory 提出了一个改进 GN 算法的重叠社区发现算法——CONGA 算法（Cluster-Overlap Newman Girvan Algorithm）。

GN 算法有两个重要的问题引起了研究者的注意，其一是该算法的时间复杂度很高，因为算法每次都要重新计算网络中所有边的边介数，当网络中的结点数目很大时，这个计算过程是非常耗时的；其二是 GN 算法是对复杂网络的一种硬分类，即每个结点有且仅能属于一个社区，这点是和实际的网络结构不相符的，因为大多数的实际网络都是有重叠结构的。

因此，CONGA 算法改进 GN 算法使之能进行重叠社区发现，其主要的贡献是：其一，定义了网络中的结点介数，结点介数是在边介数的基础上定义的，目的是找出那些结点介数高的结点，类似于边介数的意义，结点介数高的结点是重叠结点的可能性比其他结点高，若将这样的结点归属于某一个社区，这是不合理的，因此，CONGA 算法采用的办法是分裂结点介数高的结点为多个，相当于复制了一个结点的副本，原始结点和副本结点之间增加一条虚边，然后再去完成 GN 算法。其二，定义了分裂介数的概念，此概念的作用是用来判定在什么时候分裂结点以及怎样分裂结点，这一步在算法中很关键，其有机地与结点介数和边介数结合为一体促使 CONGA 算法成为完成的整体。图 3.5 解释了在网络中如何分裂结点的过程。

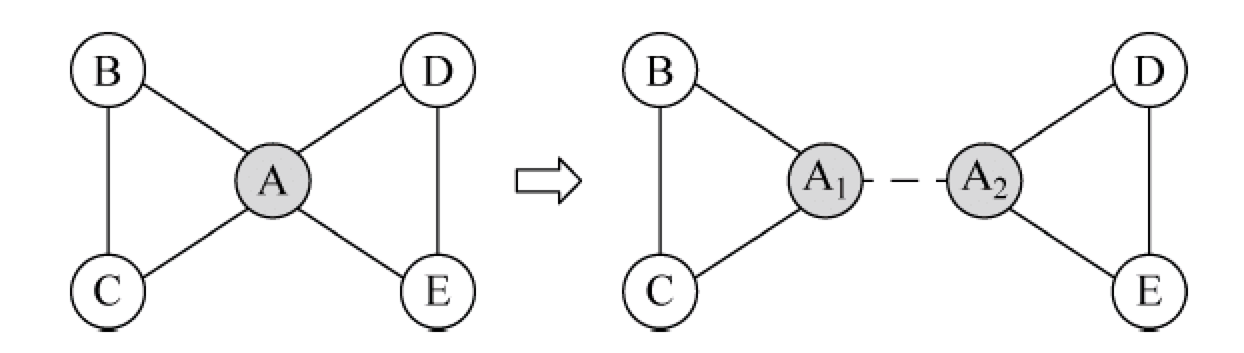


图 3.5　CONGA 算法分裂结点的示意图

（资料来源：Gregory. An Algorithm to Find Overlapping Community Structure in Networks. Proc 11th European Conf Principles & Practice of Knowledge Discovery in Databases, LNAI, 2007, 4702: 91-102.）

CONGA 算法的具体过程步骤如下所示：

（1）计算网络中所有边的边介数。

（2）利用边介数计算结点的结点介数，定义如式(3-3)所示：

$$C_B(v) = \frac{1}{2}\sum_{e\in \Gamma(v)} c_B(e) - (n-1) \tag{3-3}$$

其中，$\Gamma(v)$是以 v 为终点的边集合，n 是保留结点 v 在内的社区块中结点的个数，$C_B(e)$表示边 e 的边介数，$C_B(v)$表示结点 v 的结点介数。

(3) 构成结点介数大于最大边介数的结点候选集合。

(4) 若候选集合非空，计算候选集合中结点对的介数，以及计算候选结点的分裂介数；否则执行步骤(5)。

(5) 若最大的结点的分裂介数大于最大的边介数，则分裂这个具有最大分裂介数的结点；否则按照 GN 算法的步骤删除最大边介数的边。

(6) 删除边或分裂结点后，重新计算每个分割部分的所有剩余边的边介数。

(7) 重复步骤(2) ～ (6)，直到网络中不再有需要计算的边。

与 GN 算法相比较而言，CONGA 算法采用分裂结点的方法来发现社区之间的重叠性和重叠结点，然而，由于算法中通过分裂结点为多个复制结点，实际上是增加了网络中的结点数目，一旦网络中需要分裂的结点数目很多时，其计算过程是相当大的，即使分裂的结点数目少的情况下，CONGA 算法也只是仅仅减少了计算结点介数的时间，并没有减少算法本身的总体时间。GN 算法在最坏的时间情况下时间复杂度是 $O(m^2n)$，而 CONGA 算法的时间复杂度也是 $O(m^3)$，m 是实际网络中的边的总数目。

4. LFM 算法——基于局部扩展的算法

2009 年 Lancichinetti 提出了一种从局部出发的既可以找到重叠社区又可以找到层次结构的 LFM 算法(Local Fitness Method)。LFM 算法有两个重要的优势，第一个优势是提出了拟合度(Fitness)函数的概念，拟合度函数其实是社区定义的最直接反映。拟合度函数定义如式(3-4)所示：

$$f_G = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^\alpha} \tag{3-4}$$

其中，k_{in}^G是社区 G 内部的边数，k_{out}^G是社区 G 外部的边数，外部指的是有一个结点在社区 G 内，α 是一个控制社团规模的参数。该公式表示社区内部的边数占社区总链接边数的比重，显然，这个公式可直接类推到有权值的网络，对于有权值的网络，k_{in}^G和 k_{out}^G分别可表示社区 G 内部的权值和外部的权值。控制参数 α 的值可以用来调节社团的大小。

第二个优势是利用拟合度函数的不断优化且从局部搜索结点，从而发现网络的社区结构。作者认为在实际的复杂网络中，社区事实上基本表现出一种局部特性的状态，因此，有效的社区应从局部的自然状态开始。

LFM 算法的具体过程总结如下。

(1) 随机选择网络中的一个孤立结点(孤立结点是指未归属于任何一个社区的结点)作为社区 G 的初始成员，且初始化社区的 $k_{in}^G=0$。

(2) 计算社区 G 的所有邻居结点对 G 的拟合度函数贡献值，结点对社区的拟

合度贡献值定义如式(3-5)所示

$$f_G^a = f_{G+\{a\}} - f_{G-\{a\}} \tag{3-5}$$

其中，$f_{G+\{a\}}$ 表示社区 G 中添加结点 a 后形成的社区的拟合度函数，$f_{G-\{a\}}$ 表示社区 G 中删除结点 a 后形成的社区的拟合度函数，这个值反映了结点 a 添加到社区 G 后所引起的适应度变化。假如结点对社区的贡献值 $f_G^a>0$，意味着在社区 G 中添加结点 a 会增加社区的拟合度值，说明结点 a 应被加入到社区 G 中；反之，结点对社区的贡献值 $f_G^a<0$，意味着在社区 G 中添加结点 a 会减少社区的拟合度值，说明结点 a 应从社区 G 中删除。

(3) 选出拟合度贡献值为正值且最大的邻居结点，并将其加入到社区 G 中，得到新的社区 G'；否则，若社区 G 的所有邻居结点对 G 的拟合度贡献值都为负值时，循环过程停止，转到步骤(6)。

(4) 当社区发现变化后，即产生了新的社区 G'，重新计算社区 G' 中所有结点对社区 G' 的拟合度贡献值。假若某个结点的拟合度贡献值为负值，则将这个结点从社区 G' 中删除，得到新的社区 G''。

(5) 若步骤(4)中有结点从社区 G' 中删除了，则返回步骤(4)计算；反之，若步骤(4)中所有结点对社区 G' 的拟合度贡献值都为正值，则返回步骤(2)。

(6) 经过步骤(2) ～ (5)不断循环，算法完成了第一个随机结点所在的社区 G 的遍历搜索过程，之后，选择下一个孤立结点，返回步骤(1)继续下一个社区的探测过程，直到网络中的所有结点都已经被划分到至少一个社区为止。

LFM 算法结束后，从算法的执行过程中会看到有一些结点被划分到不止一个社区，这些结点就是重叠结点，或者说是所谓的“骑墙”结点，由此表现出社区结构的重叠性。此外，初始结点选择的随机性会给算法带来一些社区结构的不同。LFM 算法的时间复杂度主要取决于社团的大小和结点的重叠程度，若对于层次网络来说，在最坏的情况下的计算复杂度是 $O(n^2\log n)$。

5. LC 算法——基于边划分的算法

算法 1 至算法 4 也可称之为结点划分的方法，这些方法的划分思路是从结点的角度出发，把网络中的结点看作是研究对象，根据结点之间相近程度的衡量，决定了网络中的每个点的归属问题，最后得到了整个网络的划分，而且在某一时刻，每个点只能归属于唯一的一个社区，然而，在现实世界的各种复杂网络中，重叠性是显然存在的，因此，结点划分的这一特性必然给具有重叠性的复杂网络的社区发现带来困扰。

2009 年 Evans 针对这种结点划分方法的不足，提出了从边的角度出发进行社区划分的思路。以边作为研究对象而不是结点作为研究对象，这个想法是基于边在社区划分过程中在每个时刻都是属于唯一一个社区的，也就是边只能被一个社

区所包含,因此按照边之间的相近程度来对复杂网络进行社区划分,这样可避免重叠结点对划分结果的影响。Evans 首先对原始的复杂网络进行了变换,通过边表示结点,用共同相邻的结点来形成边,由此就将结点网络转换成了对应的边网络,然后在边网络中选择合适的社区划分方法,就可得到网络的社区结构。边划分的思想示意图如图 3.6 所示。

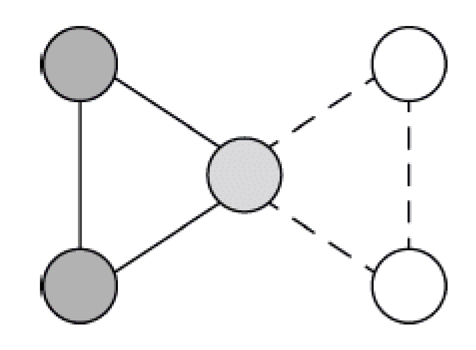


图 3.6　边社区的划分示意图

(资料来源:Evans,Lambiotte. Line Graphs,Link Partitions,and Overlapping Communities. Physical Review E,2009,80(1):016105.)

从图 3.6 中可以看出,若以边进行社区划分之后,左边实线所链接的这些结点属于一个社区,而右边虚线链接的结点属于另一个社区,社区结构是很明显的,且每条边仅属于一个社区,而浅色阴影结点就属于重叠结点,它属于两个社区,在边社区划分中,这个结点是边界结点,它并不影响社区结构的划分。

边社区划分的方法统称为 LC(Link Clustering)算法,这个算法的主要思想就是确定边之间的相似度,然后采用聚合的方法不断把相似的边聚合,或者不断把相邻且相似度高的小社区合并,反复这样的聚合过程,最后就得到整个网络的社区结构。其中,边之间的相似度的构造方法中,一种简单方法是将原有网络中的点与点的链接矩阵变换成点与边的关联矩阵,关联矩阵中包含了点与连接边之间的关系,通过矩阵变化可进一步得到边与边之间的关系矩阵,而这个关系矩阵就是边图的关系体现。另一种常用方法是利用边所链接的结点之间的链接关系来构造其相似度,如式(3-6)所示。

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \tag{3-6}$$

公式中的 $n_+(i)$表示结点 i 的广泛邻居结点,它指的是与结点 i 直接相邻接的结点所构成的集合,边 e_{ik} 的两个端点是结点 i 和结点 k,边 e_{jk} 的两个端点是结点 j 和结点 k,结点 k 是两条边的公共结点,这个公式的目的就是通过测量边的端点之间的链接共同性来反映边之间的相似性,即若结点 i 和结点 j 的链接的共同结点数目越多,或结点 i 和结点 j 的链接情况相同,则边 e_{ik} 和边 e_{jk} 的相似性越高。一旦得到了边之间的相似性的量化,就可采取聚合的方法得到以边为主体的社区划分结果,按照这种聚合的方法,通常是会得到一棵层次的树状图,在这棵层次树上选择合适的位置切割树,就可得到不一样的社区划分。有的学者在文中给出了分

区密度的判断标准,分区密度值可用来确定在树状图中进行切割的最佳位置。

边划分的社区发现方法有两个优势,一个是从结点到边的角度转变,这个转变可避免在以结点为研究目标的社区发现算法中,重叠结点处理的困扰问题;另一个是算法的复用,由于将原始网络进行转化后,对于边就可看作是硬分类问题,那么传统的基于结点的一些经典社区发现算法就可以复用在以边为研究目标的社区发现算法中。

3.3 本章小结

社会网络在变革人们生活方式的同时,也构成了 Web 上的复杂网络,以每天 TB 的量级飞速地形成了 Web 上不断变化的巨大信息源,如何利用和发现 Web 中的有用信息是企业界和研究界面临的挑战。社区是 Web 复杂网络中表现出多个个体共性的一种普遍形式,社区发现技术将为个性化服务、信息挖掘等应用提供有效的研究基础,因此,社区发现成为一个非常活跃且快速发展的研究领域。本章沿着社区发现技术的发展历程,详细展开了非重叠社区和重叠社区两大分支的发现算法分析和对比,在非重叠社区发现算法中,从最初的图分割方法、层次聚类法、分裂算法等基本算法,逐渐发展和改进,形成了包括谱方法、快速 GN 算法、基于模块度优化和基于动力学等算法。在重叠社区发现算法中,分别介绍了派系过滤算法、基于重叠模块度算法、基于局部扩展的算法和基于边划分等重叠社区发现算法。

思考题

1. 简述经典的 GN 分裂算法的思想。试设计编写相应的程序实现 GN 算法,并描述 GN 算法的优劣。

2. 试描述 2004 年由 Newman 和 Girvan 提出的模块度评价函数的概念。

3. 理解 FastNewman 算法。试设计编写相应的程序实现 FastNewman 算法,并与 GN 算法进行相应对比。

4. 简述重叠社区发现算法的目的,并选择一种重叠社区发现算法描述其思想。

5. 试从 CPM 算法、CONGA 算法、LFM 算法和 LC 算法中任选一种算法,设计并编写相应的程序实现其算法思想,并描述该算法的优劣。

6. 查找相关资料,整理社区发现算法的最新进展。

第4章

基于内容的社区聚类方法

本章学习目标

- *理解主题模型和 LDA 模型的思想和方法*
- *了解主题模型在社区发现中的应用方法*

除了前述基于网络结构的社区发现方法，基于内容的社区聚类方法也是目前社会计算领域中的研究重点。这里的内容指的就是社区内结点的文本内容。通过计算结点文本内容的相似性，能够将文本内容相似的结点划分为兴趣社区，这也就是文本聚类法。

文本相似性的度量有很多方法，主题模型(Topic Model)是文本聚类法中最典型的算法。主题模型就是对文字中隐含的主题进行建模的方法。在现实生活中，总是希望能用一种较为简单的方法来代表大规模数据集的特征信息。主题模型是一个能对大规模文本进行有效分析的模型，它不仅能够在海量互联网数据中寻找出文字间的语义主题，并且克服了传统信息检索中文档相似度计算方法的缺点。

在主题模型中，每个主题可以被表示成一个多项式的分布。如果把文本定义在文档的级别，主题模型就是抽取出文档中的语义相关的主题集合，然后将文档变换到主题空间。它能够发现文档-词项之间所蕴涵的潜在语义关系(即主题)，将文档看成一组主题的混合分布，而主题又是词语的概率分布，从而将高维度的"文档—词项"空间映射到低维度的"文档—主题—词项"空间，并捕获各个文档之间潜在的语义关系，有效提高了文本信息处理的性能。

4.1 主题模型

4.1.1 主题模型简介

在传统信息检索领域里,已经有了很多衡量文档相似性的方法,如使用向量空间模型(Vector Space Model,VSM)。传统判断两个文档相似性往往是通过比较两个文档共同出现的单词,如经典的TF-IDF(Term Frequency-Inverse Document Frequency)算法。而这些方法都是基于这样的一个假设:文档之间重复的词项越多,则这两个文档相似的可能性更大。事实上,文档的相关程度并不仅仅取决于文档本身包含的词项的重复。有很多时候,这种相关程度其实是取决于文档背后的语义关联。经常会出现这种情况,两个文档中共同出现的词项很少或是没有,但这两个文档却是相关的。例如有这样的两个句子:

“我想换个新手机。”

“不知道苹果什么时候会降价。”

虽然这两个句子中没有共同出现的词项,但当这两个句子出现在上下文时,可以很容易看出这两个句子是相关的。然而要是用传统的方法判断这两个句子肯定是不相关的,这就突显了语义关联在判断文档相关性中的重要性。如何将文档的语义关联考虑到划分中?主题模型实现了这样的功能。主题模型就是对文字中隐含主题的一种建模方法。

那么,首先确定什么是主题?主题是语料集合上语义的高度抽象和压缩表示。通俗来说,主题就是一个概念、一个方面。它表现为一系列相关的词项。进一步理解,主题就是词汇表上词项的条件概率分布。词项的条件概率越大,与主题关系越密切,反之则越疏远。在主题模型中,每个主题被表示成一个多项式分布。每个主题相对文档本身表达的内容更加抽象与压缩。

一个主题中包含了若干出现概率较高的词项。这些词项和这个主题有很强的相关性,或者说,正是这些词项共同定义了这个主题。对于一段话来说,有些词项可能属于这个主题,有些可能来自另一个主题,一段文本往往是若干个主题的杂合体。

对图4.1中的这段话,可以划分为如表4.1所示的主题。

网易做了易信,阿里力推来往,微信已经风光无二,可还是有人不服,不甘心,想要动一动腾讯在移动社交上的霸主地位。移动IM的战争早在微信打败米聊之时就已经结束了,而且,很难翻盘,因为输给微信,实乃非战之罪。

图4.1 语料示例

表 4.1　主题示例

网易	阿里	腾讯	战争
易信	来往	微信	霸主地位
移动社交	移动社交	QQ	打响
		移动社交	

可以看出，这段文字主要讲述的是其他互联网企业与腾讯微信之间移动 IM 的竞争。在这里，还出现了网易和阿里这两个主题，但它们并不是主要内容。值得注意的是，“移动社交”这样的词，既可以出现在腾讯主题，也可以出现在网易主题或是阿里主题。当它出现在具体文字中时，这三个主题都得到了一定程度的体现。

再看上面的例子，有了主题的概念后，“苹果”这个词既属于“苹果公司”这个主题，又属于“水果”的主题，如果没有语境，并不能分析出这个句子究竟属于哪个主题。但是，当联系上第一个句子时，“苹果公司”这个主题就和“手机”这个主题匹配上了，因此可以认为它们是相关的。

4.1.2　主题模型内容

通过划分为词项可以把文档表示在词项空间上。上面已经提到，主题是词项的概率分布。若指定主题模型的主题为 K 个，通过主题模型的训练，最终得到 K 个主题，就能够将词项空间中的文档变换到主题空间。主题模型是如何获得主题并表示文档的，接下来对其主要内容进行简要陈述。

1. 两个输入

主题模型处理的主体就是文档，所以将文档集合作为主题模型的一个输入。文档集合可以表示为词项—文档矩阵的形式，矩阵的内容是每个词项在每个文档中出现的次数。表 4.2 是一个词项—文档矩阵的示例。

表 4.2　词项—文档矩阵

	d_1	d_2	d_3	d_4	d_5
system	1	2	0	0	0
management	0	1	0	0	3
information	0	1	1	1	1
technology	1	1	0	0	1
intelligence	1	0	0	0	1

从表 4.2 中的矩阵可以直观地看出，一共有五个文档，每个文档又对应五个词项。文档 d_1 中 system，technology，intelligence 三个词项各出现一次。词 system 在文档 d_1 中出现一次，在文档 d_2 中出现两次。同一个词项可以在一个文档中出现多次。

除了文档集合，主题个数 K 也是一个重要输入。普遍认为，主题个数的设定是一个非常困难的问题。目前的方法大概有以下两种。

(1) 根据经验进行设定。通过反复调试或者枚举主题的数目来观察实验效果的好坏。并引入一定的评价指标进行评估。评价指标有基于困惑度(Perplexity)、语料似然值、分类正确率等。

(2) 使用非参数贝叶斯的方法。该方法先假设主题个数为无穷多个，实际主题个数随着语料规模而变化，最终能够学习出主题的数目。这个方法能在一定程度上解决主题模型中自动确定主题数目的问题，但同时也提高了运行的复杂度。

2. 可交换假设

主题模型中一个重要的假设是可交换假设，即一篇文档内的单词可以交换次序而不影响模型的训练结果。另外，文档的次序也不影响模型的训练结果。可交换也就是与顺序无关，和条件独立同分布等价。

3. 表示方法

主题模型的表示方法有两种，包括图模型和生成过程。以 LDA 模型(Latent Dirichlet Allocation)为例，使用两种方法进行表示。

图 4.2 是对 LDA 的图模型表示。

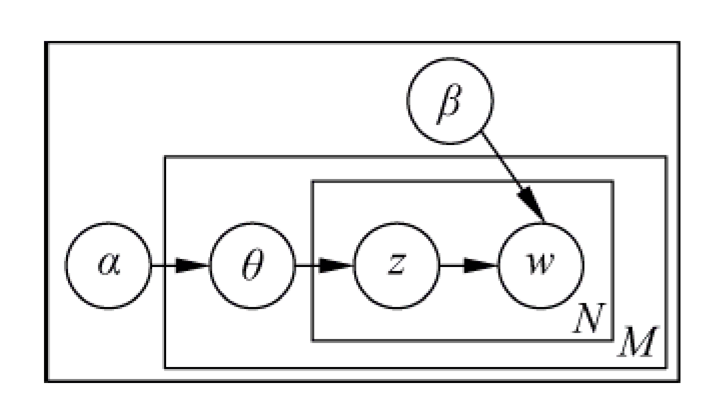


图 4.2　LDA 的图模型表示

图 4.2 中圆形代表结点，分别有观测值、隐含随机变量或参数。方框右下角的字母 $M(N)$表示方框中的内容重复 $M(N)$次。箭头代表依赖关系。其中 θ 是一个主题向量，向量的每一列表示每个主题在某文档出现的概率，文档个数为 M 个。N 表示文档长度。w 是单词，z 则是 w 所属的主题标号。α 和 β 是 Dirichlet 分布的参数。

除了图模型表示方法之外，还有一种方法来描述主题模型，那就是生成过程。LDA 模型的生成过程如图 4.3 所示。生成过程表示的是一篇文档产生的过程，若重复 M 次则能够生成整个语料集。

选择参数 $\theta \sim \mathrm{Dir}(\alpha)$
对单词 w_n
选择题目 $z_n \sim p(z|\theta)$
选择单词 $w_n \sim p(w|z)$

图 4.3　LDA 模型的生成过程表示

生成过程可以理解为，认为一篇文档的每个词都是通过以一定的概率选择了某个主题，并从这个主题中以一定的概率选择每个词这样一个过程得到的。例如，假设一个语料库中有电影、音乐、文学这三个主题。现在给定一篇介绍电影内容的文档，文档中可能同时包含了电影和音乐这两个主题。音乐这个主题中有一系列的词，这些词都与音乐有关，并且每个词分别有一个概率，代表该词在主题为音乐的文档中可能出现的频率。同样地，在电影主题中也有这样的词和概率。若想重新生成一篇关于电影内容的文档，则首先随机选择某一主题，这时选择电影和音乐这两个主题的概率更高；然后选择单词，也是选择到和两个主题相关的词的概率更高。通过不断重复这个过程，最终组成了一篇文档。当然这样得到的文档中的词是无序的。

4. 参数估计

各主题下的词项概率分布和各文档的主题概率分布是主题模型中最重要的两组参数。参数估计也就是在已知文档集的基础上，通过参数估计得到参数值的一个过程。亦即整个训练过程得到的输出结果。

5. 推断新样本

在完成对主题模型的训练后，就能够使用该主题模型推断新的样本。将表达在词项空间上的文档转换到主题空间中，得到一个以主题为坐标的低维表达。也就是得到了文档的主题概率分布。

主题模型具有灵活的扩展性，因此一经推出，就获得了广泛应用，几乎覆盖了文本挖掘和信息处理的所有领域。下面就对主题模型的几个典型类型进行介绍。

4.2　LDA 模型

早期人们处理文本、对文本进行挖掘所使用的代表方法有潜在语义分析(Latent Semantic Analysis，LSA)。潜在语义分析打破了人们以往认为文本是表示在词典空间上的思维定势。它引入了语义维度，使得文本表示从文档→词变成了文档→语义→词。然后通过线性代数的方法提取出语义维度并实现降维。在此基础上，Hofmann 提出了概率潜在语义索引(probabilistic LSI，pLSI)。而 Blei 等人在 pLSI 基础上进行扩展，于 2003 年提出了 LDA 模型。LDA 模型是目前主题

模型中应用最为广泛的一种。

4.2.1 LDA 模型简介

pLSI 寻找一个从词项空间到隐性语义(即主题)空间的变换,但 pLSI 是一个概率生成模型,而且选择了不同的最优化目标函数。LDA 模型是在 pLSI 的基础上,用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题概率分布,模拟文档的产生过程。LDA 模型其实是一种分层贝叶斯框架下的概率模型。

LDA 是一种非监督学习技术,可以用来识别大规模文档集或语料库中潜藏的主题信息。通过将文档表示为一个主题向量而不是词项向量来达到特征降维的目的。它采用的是词袋(Bag of Words)的方法。词袋方法具体来说就是将每篇文档看作一个词频向量,以此将文本信息转化为数字信息。如果直接基于词袋在文档空间对文档进行表示,会导致维度较大。若指定主题个数为 K 个,通过训练和推理,得到 K 个主题,则可以将文档变换到主题空间,从而实现降维。另外,词袋方法忽视了词与词之间的顺序,将问题简单化,同时也为模型的改进提供了契机。LDA 的基本思想与大多数主题模型思想保持一致:将文档表示为潜在主题的随机混合,其中每个主题由单词的一个概率分布来描述。

4.2.2 LDA 模型内容

传统的文档表示方法一般将文档表示为特征向量。通过使用 LDA,文档矩阵能够产生两种不同的矩阵:文档—主题矩阵和主题—词项矩阵,

$$\begin{array}{c} \quad d \\ w\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \end{array} = \begin{array}{c} \quad t \\ w\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \end{array} \times \begin{array}{c} \quad d \\ t\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \end{array}$$

其中 d 表示文档,w 表示词,t 表示主题。

由前已知 LDA 模型的框架。LDA 模型可以随机生成一篇由 N 个主题组成的文档。可以使用图 4.4 的方法来生成文档。具体描述为:

(1) 选择该文档的长度为 N。

(2) 从具有参数 α 的 Dirichlet 分布中选择一个多项式分布 θ,θ 表示每个主题发生的概率。

(3) 根据给定的 θ 确定主题 z。

(4) 根据给定的主题 z 的概率分布选择单词 w。

(5) 重复(2)和(3)N 次,直到生成全部 N 个词。

对每篇文档 d
选择参数 $\theta \sim \mathrm{Dir}(\alpha)$
对单词 w_n
选择题目 $z_n \sim p(z|\theta)$
选择单词 $w_n \sim p(w|z)$
对每一对文档 d, d'
画一个双向连接显示函数
$y|d, d \sim \Psi(\cdot|z_d, z_d)$

图 4.4　LDA 文档生成过程

其中 z_n 表示选择的主题，$p(z|\theta)$ 表示给定 θ 时主题 z 的概率分布，具体即为 θ 的值。θ 服从 Dirichlet 分布，该分布函数如式(4-1)所示：

$$\mathrm{Dir}(\mu \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \prod_{k-1}^{K} \mu_k^{\alpha_k-1} \tag{4-1}$$

其中，$0 \leqslant \mu_k \leqslant 1$，$\sum_k \mu_k = 1$；$\alpha_0 = \sum_{k-1}^{K} \alpha_k$，$\Gamma$ 是伽马函数。

这种方法首先选定一个主题向量 θ，确定每个主题被选择的概率。然后从主题分布向量 θ 中选择一个主题 z，按照主题 z 的单词概率分布生成一个单词。根据图 4.5 所示进行循环生成单词。由此可得 LDA 的联合概率如式(4-2)所示：

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{4-2}$$

上式中，N 表示文档长度；θ 表示文档的主题概率分布；w_n 表示文档的第 n 个单词；z_n 表示 w_n 所属的主题。这个函数使用 α 和 β 作为参数，通过对目标函数进行最大化来估计 α 和 β 的值。由于两个参数的耦合，它们无法直接计算出来，因此通常考虑词汇对于主题的后验概率。

按照图 4.5 所示的关系来理解，可以看出，α 和 β 表示语料级别的参数，即每个文档都一样，因此生成过程只采样一次；θ 是文档级别的变量，每个文档都对应一

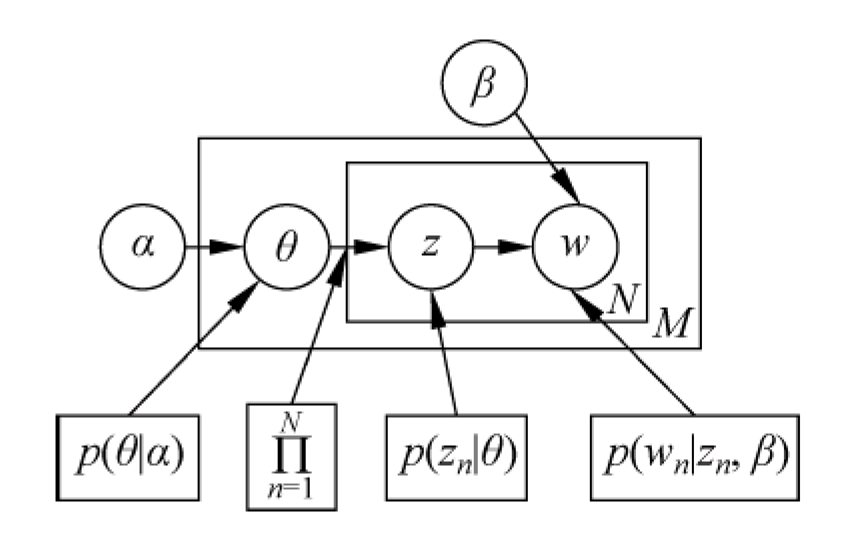


图 4.5　LDA 图模型的具体阐释

(资料来源：Blei，Ng，Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research，2003，3：993-1022.)

个 θ,每个文档产生各个主题的概率并不相同,因此生成过程中每个文档采样一次;z 和 w 都是单词级别的变量,z 由 θ 生成,w 由 z 和 β 共同生成,一个单词 w 对应一个主题 z。也就是说,LDA 模型是从给定的输入语料集中学习并训练两个参数 α 和 β,训练得到这两个参数后就确定了模型,由此可以生成文档。

4.2.3 LDA 模型统计推断

目前用来估计主题模型参数的算法有很多,如 EM(Expectation-Maximization)算法,近似推断方法的变种 EM,期望增值,Gibbs 采样等。

在 Dirichlet 先验知识和允许从后验分布的局部最大化中进行联合估计的前提下,Gibbs 采样方法最为简便和有效。Gibbs 采样方法提供了一个简单地在 Dirichlet 优先下获得参数估计的方法,并且允许这些来自于很多后验分布的局部最大值的估计进行组合。Gibbs 采样算法可认为从一个已知的主题模型中生成人造的文档数据,同时用这个算法来检查它是否可以推断原始的生成结构。它直接去估计 z 的后验分布,即每个词项到主题的分配,而不是直接估计主题—词项的分布信息和每篇文章的主题分布信息 θ。

4.3 LDA 模型的变形

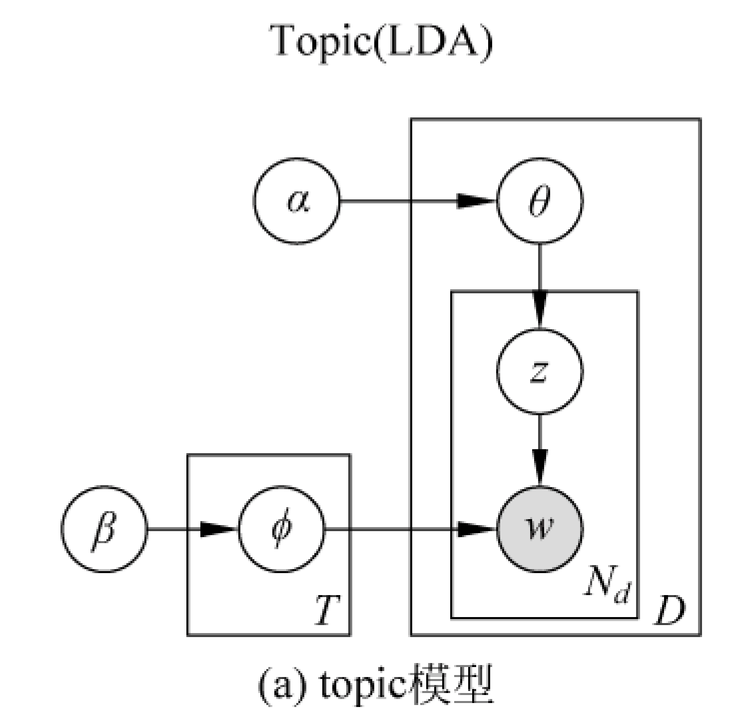
LDA 模型在推出后得到了各种各样的应用,大批的学者对 LDA 模型进行了各种变形和拓展。目前与主题模型相关的工作有很大一部分是对 LDA 模型进行修改,或是将 LDA 模型作为整个概率模型的一个部件。针对 LDA 扩展的研究工作非常多,难以对其进行全面介绍。这里只对经常被用于从社会计算角度扩展的模型进行代表性介绍。

4.3.1 AT 模型

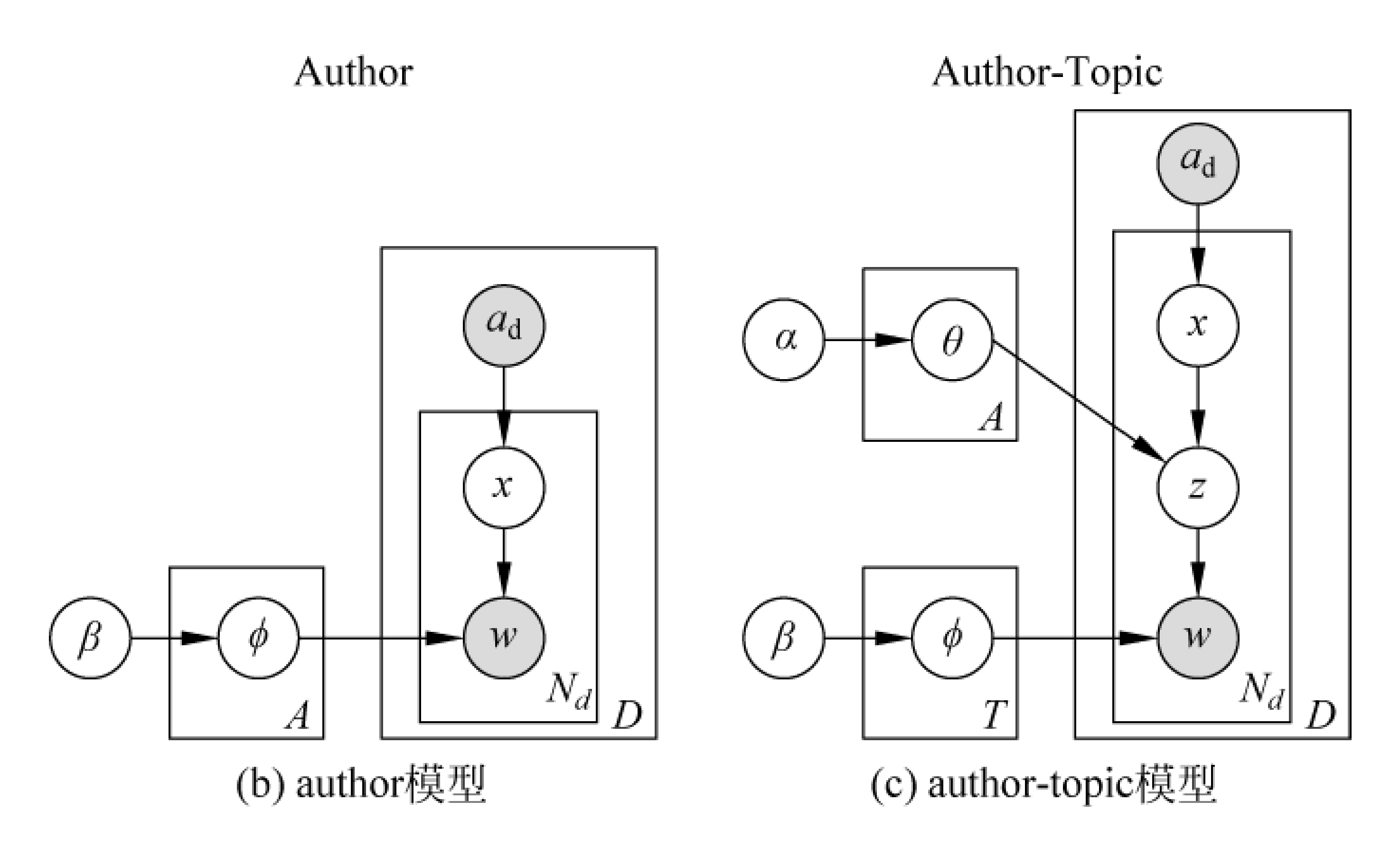
Syeyvers 等(2004)提出了 AT(Author-Topic)模型用于发现用户、文档、主题和关键词之间的关系。认为主题是多个关键词的概率分布,用户也可按照某种概率分布对多个主题感兴趣。

AT 模型将文档作者引入到对文档的主题划分中,即认为每个作者有一个主题概率分布。例如某个作者关注的领域主要是数据挖掘领域,而另一个作者主要关注点是人工智能领域。一篇文档往往有多个作者,它的主题分布可以是几个作者的主题分布的一种组合。因此 AT 模型可以获得两种主题:作者的兴趣信息和文档数据的内容信息。

此时文档生成与 LDA 模型有了一些不同，过程是这样的：随机选择一个作者，根据该作者的主题概率分布生成一个词，重复该过程直到生成整个文档。具体可以由图 4.6(c)中的部分看出。主题模型对应于每个文档有单一作者的情况，作者模式对应于每个作者有单一主题的情况。通过参数，能够获得关于作者典型写作的主题和从这些主题出发的每一个文档内容的表现方法。



(a) topic模型



(b) author模型　(c) author-topic模型

图 4.6　模型展示

(资料来源：Rosen-Zvi，Griffiths，Steyvers. The Author-Topic Model for Authors and Documents. Proc 20th Conf Uncertainty in Artificial Intelligence，AUAI Press，2004：487-494.)

4.3.2　ART 模型

McCallum 等(2007)基于发送—接收关系提出了 ART(Author-Recipient-Topic)模型，用于聚类具有相似兴趣的用户。这个模型针对的是具有方向性的文档，如电子邮件。它除了可以发现文档的内容信息之外，还可以挖掘发送者和接收者的关系。

将发送者和接收者看成是一篇文档的主题概率分布的决定因素。通过积分或

求和可以分别得到同一个人在发送者和接受者这两个角色时的主题概率分布。然后，通过使用这些主题概率分布进行聚类，判定哪些人具有相似的社会角色。例如，若有些人作为接收者时总是收到诸如要求复印、旅行预约或是安排会议室等信息，那么公认为他们具有“行政助理”这样的社会角色，即使这些人所处的社会关系完全不同。

ART 模型的生成过程是这样的：随机选择一个用户，随机确定用户的身份是发送者还是接收者，根据用户在这个身份下的主题概率分布生成一个词，重复该过程直到生成整个文档。具体过程可以由图 4.7(d)中的部分看出。

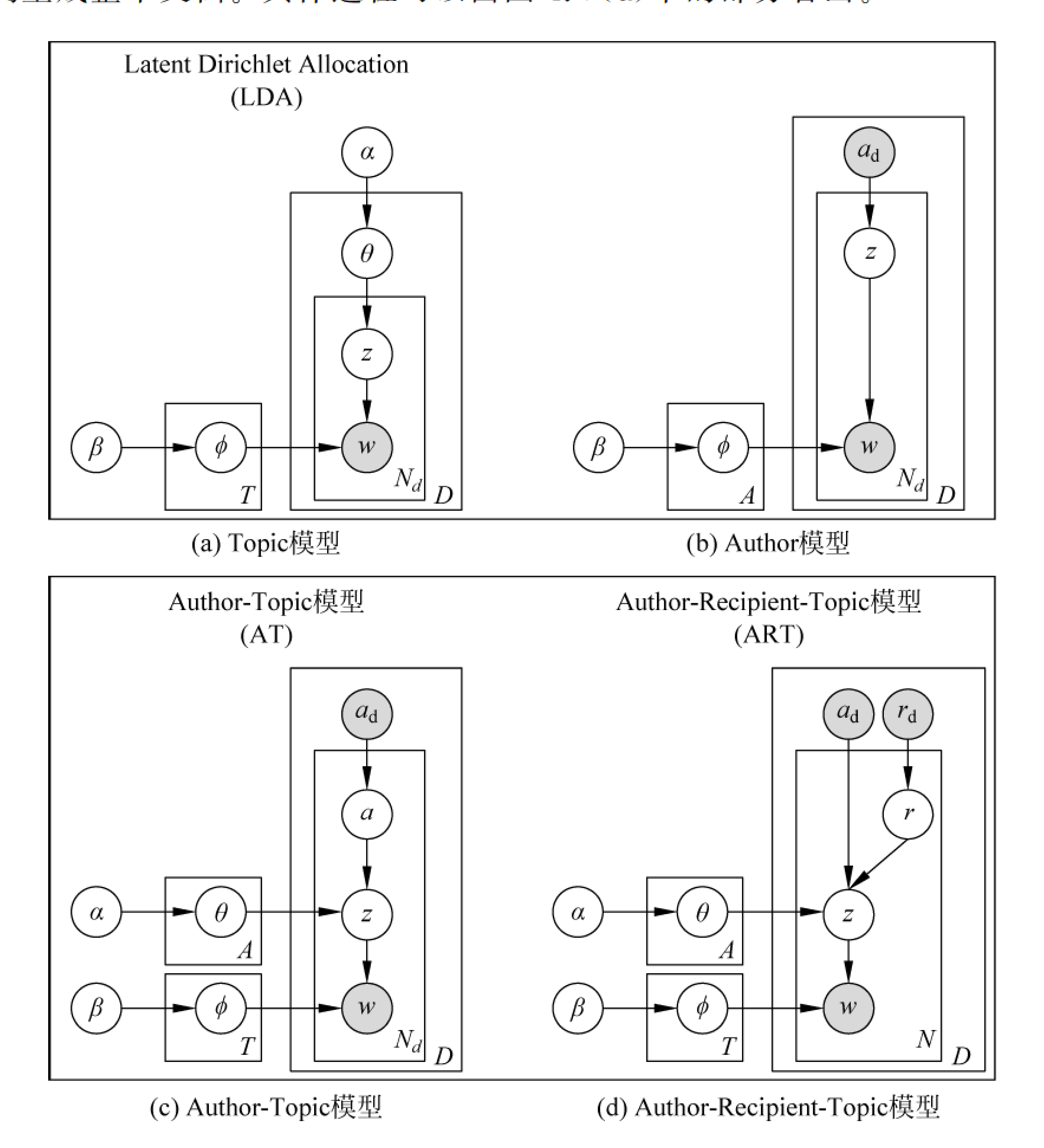


图 4.7 ART 模型展示

（资料来源：McCallum，Wang，Corrada-Emmanuel. Topic and Role Discovery in Social Networks With Experiments on Enron and Academic Email. J Artificial Intelligence Research，2007，30：249-272.）

4.3.3　CART 模型

在 ART 模型的基础上，Pathak 等(2008)提出了 CART(Community-Author-Recipient-Topic)模型。该模型适用于提取电子邮件网络中隐含的子社区，但仅适用于有一个发送者和多个接收者的有向网络，因此不适用于研究如 BBS 论坛或是这种存在用户双向交流的社会网络中的社区发现。

4.4　主题模型在社区发现中的应用

4.4.1　简介

传统的基于结构的社会网络分析方法中，大多数研究通过分析用户之间链接行为以发现重要结点和社区演化特征等，但是这样的效果并不理想。研究兴趣或领域的相近关系，才是反映社区性质的最重要的关系。随着研究的深入，研究者们开始关注社会网络信息中更丰富的内容信息。信息内容的加入有助于为社会网络潜在问题的研究创造更多条件。

在传统的社区结构分析方法中，往往忽视了结点内容，没有考虑从内容层面形成的链接关系或是由于线下活动而形成的潜在的链接关系。区别于传统的社区发现方法，基于主题模型的社区发现方法，可以通过引入主题模型获得主题信息，构建关联网络，在此基础上进行的社区发现研究，更加符合现实的需求并赋予社区一定的意义。

4.4.2　网络结构挖掘

网络结构是文本的一个附加信息，存在于各种类型的数据集合中：如微博中的好友关系、网页中的链接关系等。这些链接可以更好地帮助分析文档的语义含义，而文档的语义含义也可以更好地帮助分析网络结构的链接关系。对于一些数据类型来说，例如社会关系的网络数据，链接关系本身就是一种数据类型，处于和文本同样重要的地位。接下来介绍一些常见的将主题模型融入网络结构信息的方法。

1. 利用相似性

首先介绍 RTM(Relational Topic Model)模型，与标准的 LDA 模型相比，它多了第二个步骤，试图从主题分布的相似性来考虑进一步生成附加的链接结构。这是包含了一个隐含的假设：如果两个文档之间有着链接关系，那么它们之间的主

题分布应该更为相似。

在给定文档的主题分布后,使用链接生成函数生成链接关系。除此之外,还可以在最后的链接关系形成函数中引入社会关系。也就是说,两个结点形成链接关系是由于两个因素:(1)主题分布的相似性;(2)社区从属关系的相似性。另外一种方法是,考虑每次链接事件的形成过程:首先为整个链接事件选择一个主题标签;然后基于该主题为参与到该事件的两个结点进行社区标签采样;最后基于这两个社区标签决定一个链接事件形成的概率。

2. 利用规则化方法

规则化的基本思想是在模型的最优化函数上添加一些限制,通过这些限制使得模型避免出现过度拟合等病态学习问题。对于大多数主题模型,如标准的 LDA 和 pLSA 的求解过程都是一个最优化的过程,目标函数就是使得该模型在语料集合上的似然最大。在主题模型中可以加入网络规则化因子。基本思想与上述相同:如果两个结点存在链接关系,那么它们存在相似性。例如,如果两个博主转发过同一条微博,那么他们的主题兴趣分布应该比较相似。

3. 隐式聚类

上述方法都是显式地利用网络结构特征进行聚类形成网络子结构的方法。还有一种方法是不显式地对链接进行建模。ATM 就可以看成一种隐式聚类的方法,亦即可以按照主题的分布对作者进行聚类,例如将作者归到他具有最大数值的主题内的群体。基于 ATM 提出过这样的模型,在主题模型中增加一个子结构变量(例如,在社会关系网络中,社区就是子结构),所有结点都和这些子结构建立联系,形成一种类似星形的结构,这与前两种方法有着本质上的不同。基本思想是在 ATM 的基础上引入社区的变量,然后设定每个社区在作者集合上有一个多项式分布。

4.5 本章小结

现实世界中的许多复杂系统都在不同程度上体现出社区的特性,单纯从网络链接结构来划分社会网络的社区结构是不够的,社会网络中的内容信息也是不容忽视的重要部分。本章首先从文本挖掘的角度出发,详细介绍了常用的主题模型、潜在语义的概率模型以及不断改进的各种信息融合的语义概率模型;最后,本章描述了如何在社区发现技术中应用主题模型,以及主题模型融入网络结构信息的方法。

思考题

1. 简述主题模型的思想，并根据你的理解，试描述主题模型如何应用到社会网络中进行社区发现的设想。

2. 简述潜在语义模型的思想，根据你的理解，试描述潜在语义模型如何应用到社会网络中进行社区发现的设想。

3. 查找相关资料，分析社会网络中内容信息如何应用到社区发现技术中的最新进展。

第5章

社会网络信息传播分析

本章学习目标

- 理解社会网络中的信息传播的意义和概念
- 熟悉和理解经典的社会网络的信息传播模型
- 了解社会网络中的信息传播的应用

5.1 社会网络中的信息传播

社会网络(Social Networks)可追溯于哈佛大学的著名社会心理学家米尔格伦(Stanley Milgram)在20世纪60年代(Milgram,1967)所证明的六度分隔理论(Six Degrees of Separation),也被称为小世界理论(Easley 和 Kleinberg,2010)。该理论通过一个连锁信件实验精妙地说明:任何两个陌生人之间所间隔的人不会超过六个,也就是说,任何一个陌生人最多通过六个人就能够认识另一个陌生人。随着互联网技术的迅猛发展和现代生活模式的改变,各种社交网络不断涌现,这些网络为人们提供了社会化的网络服务,如国外的Facebook、Twitter、Flicker,国内的腾讯QQ、新浪微博等,这些平台为用户提供了一个以关系链接和信息生产和分享为主的社交服务应用,用户之间通过相互之间的朋友或关注关系形成了类似于现实社会中的用户关系的虚拟社会网络世界。在这样的虚拟社会平台上,每个用户个体通过生产和分享各自的信息内容,使得大量的信息沿着用户之间的关系链条(或说联系路径)不断地传播开来。

社会网络也可以说是社会化媒体的结构表现形式,表5.1中总结了社会化媒体的各种类型(Tang 和 Liu,2012),从中可以发现,人们已经无时无刻地与社会化媒体产生着联系,人们的生活和工作与社会化媒体有着千丝万缕的关系,同样,我

们也感受到社会化媒体所带给人们的生活方式的改变。

表 5.1 社会化媒体的类型展示

博客	Wordpress、Blogspot、LiveJournal、BlogCatalog、新浪博客、网易博客
论坛	Yahoo! answers、Epinions、大众点评网
媒体共享平台	Flickr、YouTube、Justin. tv、Ustream、Scribd、优酷
微博	Twitter、Foursquare、Google buzz、新浪微博、腾讯微博
社会网络	Facebook、MySpace、LinkedIn、Orkut、PatientsLikeMe、人人网
社会新闻	Digg、Reddit
社会标记	Del. icio. us、StumbleUpon、Diigo、QQ 书签
维基百科	Wikipedia、Scholarpedia、AskDrWiki、百度百科

社会网络中的传播现象也引起了研究者的浓厚兴趣，因为这种网络现象与真实世界中人们之间的关系和传播方式是相互对应的，只不过表现的形式各不相同而已。社会网络中的传播包括新闻的传播、观点和想法的传播、新产品的采用、潮流的引领等，在生活中，社会网络也有许多应用实例，如纽约电力网格、朋友圈网络、斑马群网络等，而社会网络的影响传播与现实生活也是息息相关的，且以一种显著的令人称奇的方式缩短着世界的距离，跨越了国界，跨越了时空。因此，在现实生活中越来越深刻感觉到社会化媒体所带来的力量，正像网络中描述的那样，社会化媒体的好处在于其能够使得虚拟的世界和真实的世界进行交汇。举一些社会网络传播所带来效应的例子，例如：2012 年一位妻子在网络上发帖，称她的丈夫肾衰竭，希望能找到 O 型捐助者，网络将此信息不断地传播开来，一名喜剧演员予以响应并且正好配型吻合，移植手术非常成功。再如，2012 年在美国总统竞选中，奥巴马通过对新媒体潜能的充分发掘，颠覆了传统的竞选方式，社会化媒体让选民比任何时候都更充分地参与到大选之中，并且参选人也以尽可能少的成本让尽可能多的公民相信他并给他投票，这就是社会化媒体所发挥的传播作用所产生的结果。

事实上，信息是个有广泛含义的概念，指一切可以被人类所能接收到的东西，同时，信息的意义也是需要通过传播来体现。信息传播是人与人之间交换信息的行为，它在人们的日常生活中无处不在，是人们生活中的重要组成部分，信息传播可以促进人们对世界及彼此的了解。在知识经济时代，信息传播显得尤为重要，掌握信息优势已经成为赢得竞争的重要前提条件。过去，普通用户作为信息的消费者，从电视、收音机、电影和报纸等传统媒体上观察世界。然而随着互联网的发展与 Web 2.0 时代的到来，用户不仅仅是信息的消费者，同时也是信息的创造者和传播者。近年来涌现出的万维社会化媒体成为信息传播的重要平台，用户与用户之间及用户与媒体之间的信息传播均表现出高交互性的特征。在紧急事件发生时，这些万维社会化媒体比传统媒体反应更加迅速、灵敏、准确，并且影响广、简单

易用、高效,故研究信息传播的特征在当今互联网时代尤为重要。

5.2 社会网络中的信息传播模型

当前,对网络中的信息传播问题已进行了大量的研究,其中一部分研究的思路是借鉴病毒传播模型而改进研究的。另一部分是从影响力传播的角度出发而设计相应的模型。

5.2.1 病毒传播模型

社会网络中信息传播和病毒传播在某些方面有一定的相似之处,因此了解病毒传播模型对人们解决信息传播问题有很大的帮助。一般情况下,疾病(病毒)借用一种接触网络来进行病毒的传播,通常,网络中的每个结点代表一个人,两个结点之间的边表示人之间曾经有过接触,从而可导致疾病(病毒)就有可能从一个人(结点)传染到另一个人(结点)。疾病的传播过程有时会有突发性,有时会持续一段时间,传染的力度也和疾病的特征以及接触的人群网络有很大关系。

经典的病毒传播模型主要有 SIR 模型和 SIS 模型两种。

1. SIR 模型

著名的 SIR 病毒模型是由 Kermack 和 McKendrick(1927)在研究黑死病的传播规律时构建的。在 SIR 模型中通常是将网络中的感染对象分为三种状态,它们分别是 S 状态(称为未感染状态)、I 状态(称为传染状态)、R 状态(称为免疫状态)。未感染的个体不会感染其他的个体,但是有可能被其他的个体感染。处于传染状态的个体已经被感染,具有传染性,会感染其他的未感染个体。而免疫个体或者是已经被治愈并且获得免疫力的个体或者是已经死亡的个体,它们不会感染其他个体,也不会被其他个体感染。三种状态之间的转换关系如图 5.1 所示。

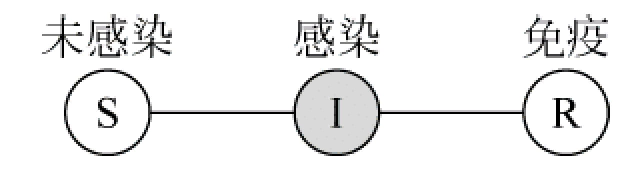


图 5.1　SIR 模型的状态变换关系图

SIR 模型适合描述那些染病者在治愈后可以获得终生免疫能力的疾病(如腮腺炎),或者是那些几乎不可避免走向死亡的疾病(如艾滋病)。

一般来说,若采用数学上的公式来表达疾病传染与时间之间的关系,通常在单位时间内,未感染的个体以平均速率 β 被感染为传染状态,又以平均速率 δ 被治愈成正常或死亡状态。此模型的数学公式如下所示。

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta I(t)S(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta I(t)S(t) - \delta I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \delta I(t) \end{cases} \tag{5-1}$$

其中，在式(5-1)中 $S(t)$、$I(t)$和 $R(t)$分别表示 t 时刻未感染个体数量、感染个体数量和被治愈的个体数量，这三个值需要满足公式的约束条件。

$$S(t)+I(t)+R(t)=N$$

下面，经过简单假设和公式变换，得到某个传染病初期的感染个体的预测。设 S_0为初始易感染人数，假定$R_0=0$，$\lambda=\dfrac{\beta}{\delta}$为病毒的传播速率，$\rho=\dfrac{\delta}{\beta}=\dfrac{1}{\lambda}$，那么，由公式之间的运算变换和变形，再经过积分变换，最终可得到式(5-2)。

$$S(t)=S_0\,\mathrm{e}^{-\frac{R(t)}{\rho}} \tag{5-2}$$

并且根据变换和变形结果，以及上式，还可得到式(5-3)。

$$\frac{dR(t)}{dt}=\delta\left[N-R(t)-S_0\,\mathrm{e}^{\left(\frac{-R(t)}{\rho}\right)}\right] \tag{5-3}$$

这个结果表达是在 t 时刻感染个体的数量，这个数量只与总量、被治愈的人数、初始未感染的人群数量有关，而这些量的获取相对来说是可以获得的，由此就可以估算出某一个时刻的感染个体的数量，如式(5-4)所示。

$$I(t)=N-R(t)-S_0\,\mathrm{e}^{\left(\frac{-R(t)}{\rho}\right)} \tag{5-4}$$

对于式(5-4)可以进行直接的求解，并且求解得到的参数具有明确的物理意义，在病毒爆发初期就可以进行预报(徐腾龙，2013)。

虽然 SIR 传播模型在许多网络中得到了扩展和研究，也是当前研究的热点，然而却不能准确地表达当前在线社交网络的传播现实，如谣言传播过程中的从众性、传播意愿的累积性等，因此根据传播关键因素建立合理的传播模型是当前研究的重点。

2. SIS 模型

与 SIR 模型不同，SIS 模型将感染对象只分为两种状态：S 状态(称为未感染状态)、I 状态(称为传染状态)。处于传染状态的个体可以通过药物或自身体质来治愈，但是在治愈后并未获得免疫能力，而是重新成为新的未感染状态的个体，并和其他的未感染个体一样，有一定的可能再次被感染。SIS 模型的基本过程是：起初接触网络的状态是，有一些处于传染状态 I 的结点和剩余处于未感染状态 S 的结点，之后，处于传染状态 I 的结点 u 在固定时间步骤内以某种概率传染给其周围处于未感染状态 S 的结点，在固定时间步骤结束后，结点 u 不再具有传染性，又一次回到未感染状态 S。因此，SIS 模型是一种在未感染状态 S 和传染状态 I 下的两

种状态的交替过程(Easley 和 Kleinberg,2010)。SIS 模型很好地描述了像流感、结核病这类无法获得免疫力的传染病。

然而,信息的传播和疾病的传播有着明显的本质上的区别。例如,信息传播活性随时间快速衰减,而疾病一般不会;信息传播中不同类型边不仅是传播力不同,传播的模式也不同,而疾病传播中接触强度只会造成传播概率差异;信息传播受到信息内容的重大影响,每次传播激活的有效网络不同等。基于以上差异,病毒传播模型在某些方面并不能较好地解释信息传播,故经过进一步的研究发现,提出了影响力传播模型。

5.2.2 影响力传播模型

经典的影响力传播的模型主要有两种:线性阈值模型(Linear Threshold Model)和独立级联模型(Independent Cascade Model)。

借助于图论的研究基础,可以把社会网络抽象成一个无向或有向图,通常采用 $G(V, E, W)$ 的形式来表示图,并且给定初始传播结点集合 S_0。其中,V 表示社会网络中的个体集合,通常集合中的个体指的是人;E 表示个体与个体之间的相互关系集合,即在无向图中存在一条边 e_{uv} 在 E 集合中,在信息传播中就表示结点 u 与结点 v 之间存在传播路径;W 则表示个体之间的关系权重,在信息传播中,w_{ij} 有时可代表个体 i 对个体 j 的影响概率,也就是结点 i 成功激活结点 j 的概率。不失一般性,假设 $\sum_{v\text{的邻接结点}u} w_{vu} \leqslant 1$。每个结点都有两种状态:未激活状态、激活状态。激活是指某个未激活状态的结点被其已激活的邻接结点所影响而变为激活状态,激活状态表示接受了某种观点或产品,未激活状态表示没有被影响到,或是指未接受观点或产品。在社会网络中,激活状态结点数目越多说明结点的影响越大。

1. 线性阈值模型

阈值模型针对的是一种集体行为,集体行为下的阈值指的是个体基于某种社会系统中已经参与某项行为活动的其他人的比例或倾向来决定是否参与该活动,由此,个体是否采纳新的行为依据的是社会系统中或是群体中其他人行为的函数。事实上,在现实生活中个体行为受群体其他人行为的影响的例子比比皆是,举个生活中常见的例子来形象地阐述阈值模型。假设当某个人去一个陌生的城市,在晚饭时间这个人想在一个不知名的餐馆吃饭,通常情况下,很多人的想法是如果在这个餐馆中的就餐的人数适中或很多,说明这个餐馆的饭的味道还不错,那么这个人就会决定在这个餐馆就餐;反之,这个人很有可能不会在这个餐馆吃饭。同样在采纳新思想或新产品时,大部分人的想法是会受其他人使用该新产品或新思想后的效果来决定是否采纳新的思想和产品。因此,这种个体接受新信息的倾向或是比例就是其接受新信息的阈值。

同样原理，在社会网络中的影响传播模型中，阈值模型就是最直接的想法。线性阈值模型的研究可以回溯到 20 世纪 70 年代，它的主要思想是：为信息传播网络中的每个结点设定一个接受阈值，结点之间都存在互相影响的可能性，这种可能性被表示成社会网络中结点与结点之间的影响概率，影响概率的值越大，表明结点之间的交互强度越大，即相互影响的强度也会越大。对于每个结点来说，随着时间时刻不断推移，如果它周围的邻居结点越来越多地接受了某种观点或实施某种共同行为，那么相邻接的结点会以各自的影响概率来影响它，一旦这些影响的概率值累积超过了这个结点的接受阈值，那么这个结点就会接受同样的观点或者实施同样的行为。由此，线性阈值模型的传播过程可以描述如下：对于每个结点 $v \in V$ 一律随机地抽取一个阈值 $\theta_v \in [0,1]$，阈值 θ_v 表示网络中的个体 v 在面对周围群体的其他行为后，是否会采取同样行为的界限，换句话说，要激活结点 v，θ_v 表示结点 v 所能接受到的已经被激活的邻接结点的累积影响强度。结点 v 所接受到的累积强度 E_u 可以表示为

$$E_v = \sum_{u\text{是}v\text{的已被激活的邻居结点}} w_{vu}$$

设定一个初始的激活结点集合 $V_0 \subset V$，其余不属于 V_0 集合的结点都是未激活的结点，激活过程是在离散的时间步长依次进行，在 $t-1$ 时刻所有被激活的活跃结点可以在 t 时刻去激活它们的未被激活的邻接结点，在 t 时刻，若某个结点 u 的累积影响强度 $E_u \geqslant \theta_u$ 时，该结点 u 就成为激活结点，那么结点 u 就具备了在下一时刻激活其邻居结点的能力，结点 u 加入到时刻 t 中激活状态的结点集合 S_t 中，继续这样的过程，直到再没有新的结点被激活，整个传播过程就停止。在不同的应用领域，对于每个结点的影响阈值的设定，有时由于缺乏评估特异性阈值的有效办法，除了一般意义上的随机选择方法之外，另一类最常用的方法是将每个结点的阈值统一设为一个定值，例如 0.5。

下面举例来说明线性阈值模型。假设该网络为有向图，共有七个结点，A、B、C、D、E、F、G 为简单起见，假定每条边上的传播概率是终点入度的倒数，即 $w_{uv} = 1/k_v$，其中 k_v 表示结点 v 的入度，每个结点的阈值设定为 0.5，假设初始的激活结点是结点 A 和 G，图 5.2 描述了在线性阈值模型条件下，网络中信息扩散的过程。如图所描述的一样，在 T_1 时刻，结点 B 由于其邻接结点 G 处于激活状态，且指向结点 B 的 $W_{GB}=1>\theta_B=0.5$，故结点 B 被激活；结点 D 由于其邻接结点 A 处于激活状态，且指向结点 D 的 $W_{AD}=1/2 \geqslant \theta_D=0.5$，故结点 D 被激活。在 T_2 时刻，结点 C 由于其邻接结点 A、B 处于激活状态，且指向结点 C 的 $W_{BC}+W_{AC}=1/3+1/3>\theta_C=0.5$，故结点 C 被激活；同理，结点 E 由于其邻接结点 B 处于激活状态，且指向结点 E 的 $W_{BE}=1/2 \geqslant \theta_E=0.5$，故结点 E 被激活；结点 F 由于其邻接结点 B、D 处于激活状态，且指向结点 F 的 $W_{BF}+W_{DF}=1/2+1/2>\theta_F=0.5$，故结点 F

被激活。至此,网络中再没有需要激活的结点时,该信息传播的过程停止。

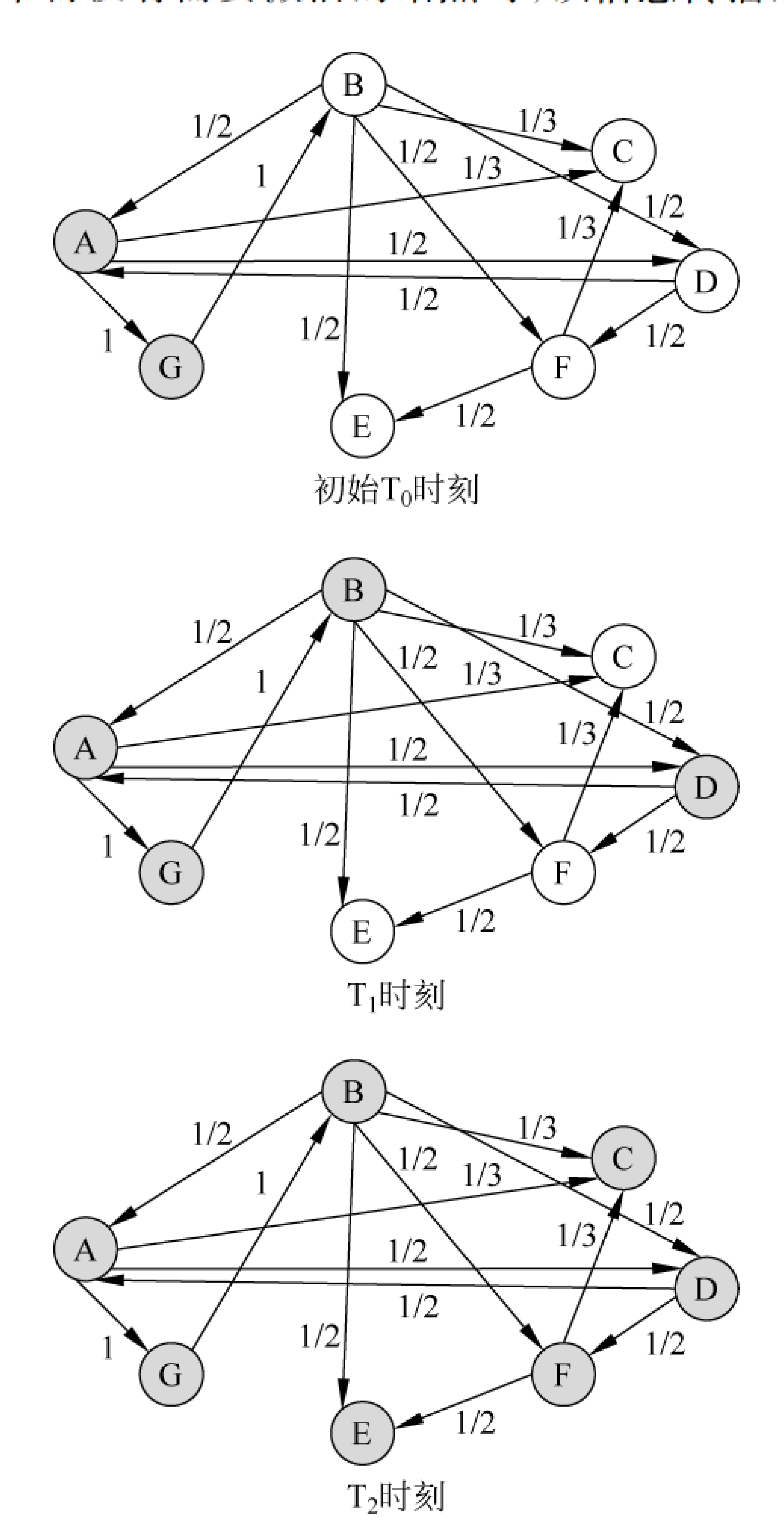


图 5.2　线性阈值模型的传播例子

2. 独立级联模型

独立级联模型也是经典的信息扩散模型,同样是一个概率模型。这个模型与线性阈值模型的关注重点是不同的,其关注的焦点是信息的发送方,即每个结点在下一时刻激活其邻居结点的机会只有一次。在独立级联模型中,信息扩散过程与线性阈值模型的信息扩散相同,都是从一个初始选定的活动结点集合作为传播源开始,一旦当某个结点 w 在第 t 步被激活,它将只有一次机会去激活它的每一个当前未被激活的邻居结点,如果对于结点 w 的某个邻居结点 v 来说,结点 v 被结点 w 激活的概率是 P_{wv},如果结点 v 被 w 成功激活,那么结点 v 就成为在第 $t+1$ 步被激活的结点,结点 v 被加入到 $t+1$ 时刻的激活结点集合,到此,结点 w 也完成了

它的激活使命，结点 w 将不能再激活它的其余的邻居结点。在这个过程中，激活过程是不可逆的，即某个未激活结点一旦被激活，便再无法改变其变为未激活结点。对于结点集合中的每个结点重复刚才的过程，直至没有新的结点被激活时，这个传播过程就停止。

下面举例来说明独立级联模型。假设该网络为有向图，如图 5.2 一样共有七个结点，A、B、C、D、E、F、G，每条边的权值表示结点 u 以 P_{uv} 的概率激活 v，其中 u 表示边的起点，v 表示边的终点，为了简单起见，随机给定每条边上的传播概率，图 5.3 描述了在独立级联模型下，网络中信息扩散的过程。假设初始的激活结点仍然是结点 A 和 G，如图所描述的一样，在 T_1 时刻，结点 G 试图激活其邻接结点 B，结点 A 试图激活其邻接结点 C、D，在这种情况下，假设只有结点 B 被成功激活，结点 C 和 D 都被激活失败，由此，结点 A 和 G 完成了它们的激活使命，在下一时刻 T_2 结点 A 和 G 就不再有机会激活其他结点；在 T_2 时刻，结点 B 分别以不同的概率试图激活其邻接结点 C、E、F 和 D，但是，只有结点 E、F 被成功激活，结点 C 和 D 激活失败，由此，结点 B 在时刻 T_2 完成了它的激活使命，在下一时刻 T_3 结点 B 就不再有机会激活其他结点；在 T_3 时刻，结点 E 和 F 分别以不同的概率试图激活其邻接结点 C，此时，假设结点 C 被成功激活，则结点 E 和 F 的激活使命终结，由于在下一时刻 C 不能影响任何结点，到此，所有结点都各自完成了它们的激活任务，则网络中的信息扩散过程全部结束。

3. 其他模型

关于社会网络中的信息传播，还有许多其他的模型，例如有竞争的影响传播模型，有负面评价的传播模型，观点舆论传播模型，投票模型等。

对于一个有争议的舆论，一个人可以选择支持或反对。然而在观点传播过程中，个体可能由于信任程度、周围邻居鼓动等对固有观点产生动摇，从而对全局观点的分布产生影响。当一个群体达成共识时，才能发挥集体的力量。

最经典的观点舆论模型是 Sznajd-weron 和 Sznajd 在 2000 年提出的基于类自旋系统模型。该模型以投票选举 A 和 B 为例，每个人都会按照自己的观点来选择支持 A 或 B，并且为网络中的结点建立一维规则的格子链，每个格子表示一个个体。每个个体的观点 $S_i=1$ 表示赞同，$S_i=-1$ 表示反对。

观点在网络中的传播规则如下：

（1）如果 $S_iS_{i+1}=1$，则 S_{i-1} 和 S_{i+2} 都和 S_i，S_{i+1} 取值相同。

（2）如果 $S_iS_{i+1}=-1$，则 S_{i-1} 和 S_{i+2} 分别取 S_{i+1} 和 S_i 的数值。

这个规则阐述了这样的思想：一对结点的观点可以影响其共同邻居的观点。即当一对结点有着相同的见解，它们最邻近的共同邻居就持相同观点。若当一对结点持不同观点，其最近的邻居的每一个成员意见都不一样。

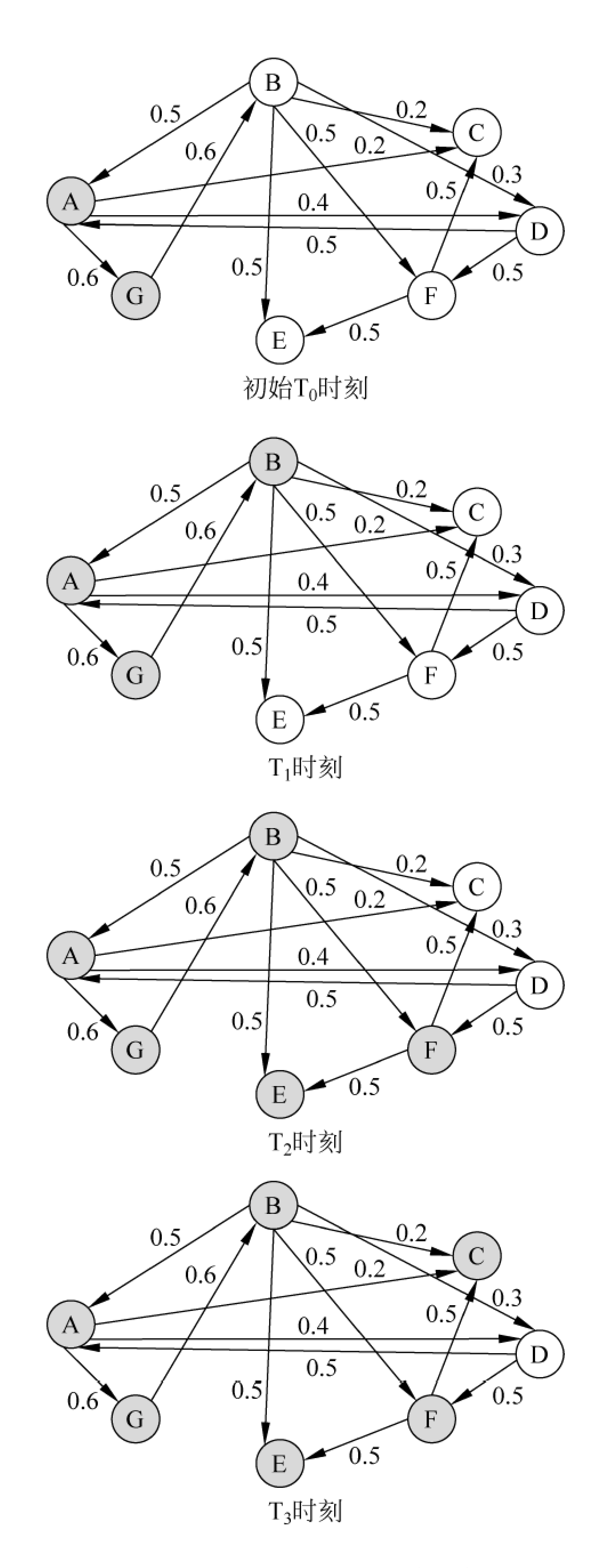


图 5.3 独立级联模型的传播例子

利用该模型分别研究了所有人支持 A，所有人支持 B，50%支持 A 和 50%支持 B 的三种情况下观点的磁化率以及信息噪声。结果表明在一个封闭的社会，一

种观点要么出现独裁要么就是僵局，总体趋于观点一致。而在小而开放的社会，观点不会趋向于任何稳定状态（孟繁荣，2013）。

5.3　社会网络中的信息传播的应用

5.3.1　影响最大化

由于社会网络是由个体及个体之间的关系所组成的一个复杂网络，这种复杂的社会结构对信息的传播和扩散起着至关重要的作用。例如当一个人采纳一个新的思想或接受一种产品时，他会向他的朋友或同事推荐，某些人可能会接受或采纳他的推荐，并进一步向他们自己的朋友或同事推荐，一个人的行为在很大程度上取决于周边的朋友或同事的决定。而影响最大化问题主要考虑了如何有效地发挥一个人在社会网络中的影响力。

影响最大化问题的研究有着十分重要的现实意义，在市场营销、广告发布、舆情预警以及社会安定等方面有十分重要的应用。影响最大化的问题实际上可归结为这样一个问题：即如何在网络中选择一些初始的受众群体，让他们接受某种产品和新思想，然后再利用他们的影响力，不断地把这种产品的效果和思想传播出去，这就是口口相传的目的，为了在尽可能少的成本下使得被影响的范围扩大，即影响广大群体。一个关键性的问题就是如何选择这些初始的受众，使得选择的人又少，但其影响的面又很广。

Richardson 等（2002）将影响最大化问题归纳为一个算法问题。近年来，社会网络中影响最大化算法成为研究热点。一部分研究的关注点在于寻找网络中最有影响力的那些结点；一部分研究的目标主要集中在如何扩大影响范围同时降低算法的时间复杂度；还有的研究是基于在不同情况下的影响最大化问题，例如有负面信息存在的情况下，有竞争信息存在的情况下。当前，社会网络影响最大化问题的研究都会基于之前所介绍的两个基本传播模型：线性阈值模型和独立级联模型。

随着 Web 2.0 的出现及流行，目前出现了很多大型在线社交网站，如上文提到的 Facebook、Flickr 等，这些大型在线社会网络的成员数目都非常庞大，数据量的极大增加对传统社会网络中的影响最大化算法，包括传播模型均提出了巨大的挑战。

5.3.2　病毒营销

病毒营销（Viral Marketing，又称病毒式营销、病毒性营销、基因行销或核爆式行销），是一种常用的网络营销方法，常用于进行网站推广、品牌推广等。病毒营销

利用的是用户口碑传播的原理，在互联网上，这种“口碑传播”更为方便，可以像病毒一样迅速蔓延，利用快速复制的方式传向数以千计、数以百万计的受众，因此病毒营销成为一种高效的信息传播方式。而且，由于这种传播是用户之间自发进行的，因此是一种几乎不需要费用的网络营销手段，并能迅速地扩大自己的影响。

病毒营销这一概念，最早由贾维逊(Steve Jurvetson)及德雷伯(Tim Draper)在 1997 年发表的《病毒营销》一文中首先提出，并初步定义为“基于网络的口碑传播”。这个概念的提出是基于 Hotmail 的实践。Hotmail 是世界上最大的免费电子邮件服务提供商，在创建之后的 1 年半时间里，就取得了令人不可思议的成绩。它吸引了 1200 万注册用户，而且还在以每天超过 15 万新用户的速度发展。当时 Hotmail.com 提供免费 E-mail 地址和服务，在每一封免费发出的信息底部附加一个简单标签：Get your private, free email at http://www.hotmail.com。人们可以利用免费的 E-mail 向朋友或同事发送信息，并且接收邮件的人也将看到邮件底部的信息，同时应邀加入使用免费 E-mail 服务的行列。通过上述的策略，Hotmail 提供免费 E-mail 的信息在更大的范围扩散。而同期的竞争者 Juno Online Services 没有采用病毒式营销的推广方式，而是在传统的营销方式上斥资 2000 万美元，但收效甚微。由此可见病毒营销的效果和威力。

Hotmail 之所以爆炸式的发展，就是由于利用了“病毒式营销”的巨大效力。病毒式营销的成功案例还包括 Amazon、ICQ、eGroups 等国际著名网络公司。病毒式营销既可以被看作是一种网络营销方法，也可以被认为是一种网络营销思想，即通过提供有价值的信息和服务，利用用户之间的主动传播来实现网络营销信息传递的目的。

病毒营销主要有以下特点：

(1) 有吸引力的病原体。病毒营销能够让目标消费者自发地成为其信息传播渠道的原因在于第一传播者传递给目标群的信息不是赤裸裸的广告信息，而是经过加工的、具有很大吸引力的产品和品牌信息，这种方法为广告信息披上了一件漂亮的外衣。这使得消费者能够克服戒备心理的“防火墙”，积极接受信息，完成从纯粹受众到积极传播者的变化。

(2) 几何倍数的传播速度。病毒营销师的信息推广方式是自发的、具有扩张性的，而不是均衡地、同时地、无分别地传给社会上每一个人，这使得信息的传递更具有针对性和强渗透能力。目标受众会把信息传递给更适合接受的周围的好友、同事，从而无数个参与的“转发大军”就构成了成几何倍数传播的主力。

(3) 高效率的接收。由于在病毒营销中，信息是受众从熟悉的人那里获得或是主动搜索而来的，在接受过程中自然会有积极的心态；接收渠道也比较私人化，如手机短信、电子邮件、封闭论坛等(存在几个人同时阅读的情况，这样反而扩大了

传播效果)。以上几方面的优势,使得病毒式营销尽可能大地克服了大众媒体信息传播中的缺陷,增强了传播的效果。

(4) 更新速度快。病毒式营销的传播过程通常是呈 S 形曲线的,即在开始时很慢,当其扩大至受众的一半时速度加快,而接近最大饱和点时又慢下来。

5.3.3　谣言的防控

网络谣言是指通过网络介质(例如邮箱、聊天软件、社交网站、网络论坛等)而传播的没有事实依据的话语。主要涉及突发事件、公共领域、名人要员、颠覆传统、离经叛道等内容。谣言传播具有突发性且流传速度极快,因此对正常的社会秩序易造成不良影响。例如 2012 年上半年出现的加碘盐可以防核辐射的谣言,就导致了多个城市的抢购潮,虽然谣言很快被平息,却造成了很大的社会影响。

由于谣言在社会网络中的散布和病毒扩散很相似,Daley 和 Kendan 于 20 世纪 60 年代借鉴传染病模型提出了谣言传播的数学模型(DK 模型),在谣言传播的定量研究中被广泛地运用。此后,许多学者为扩展其应用范围,构建了形形色色的 DK 改进模型(如 MT 模型)。另外,又有学者在谣言传播模型中增加了度关联函数,并对谣言传播的复杂性、心理特征、蝴蝶效应进行深入思考等(王长春和陈超,2012)。

5.4　本章小结

社会网络中的信息传播与现实社会中的人际之间的传播行为一样,对人类的生活会产生重要的社会和经济影响,因此,社会网络中的信息传播分析研究具有重要的意义。本章首先从社会网络中信息传播的模型入手,介绍了 SIR 和 SIS 两种经典的病毒传播模型,病毒传播过程与网络中的信息传播有着非常相似的传播机理;接着,介绍了线性阈值和独立级联两种经典的影响力传播模型,以及经典的观点舆论模型;最后介绍了社会网络中的信息传播在病毒营销、谣言防控等领域的主要应用,阐述了各自的应用状况和特点。

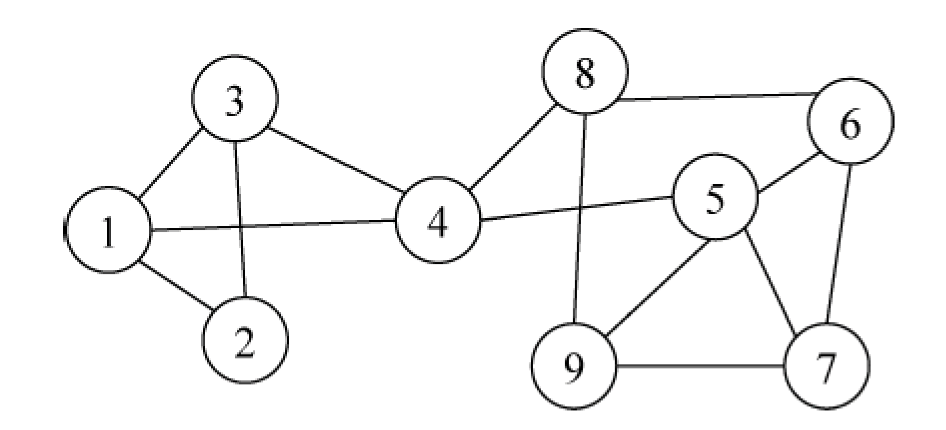
思考题

1. 简述经典的病毒传播模型 SIS 和 SIR 模型的原理。

2. 查找相关资料,分析传染病模型的最新发展状况。

3. 假设如下图所示的社会网络,假定每条边的影响权值是有方向的,即对于边 $E(u,v)$来说,W_{uv} 与 W_{vu} 是不同的,不妨假设每条边的影响权值分别是$W_{uv}=1/k_v$,

$W_{vu}=1/k_u$，其中k_v表示结点 v 的出度。设每个结点的阈值都为 0.5，设定初始的活动（激活）结点是结点 6 和 7，试画出在线性阈值下该社会网络信息传播的过程。



4. 描述下独立阈值模型与线性阈值模型的区别。

5. 查找相关资料，分析影响力传播模型的发展现状，针对有竞争的传播模型、有负面影响的传播模型，试分析下它们各自的问题和解决方法。

6. 试着搭建一个影响力传播模型的实验平台，选择常用的标准数据集，如合作者网络、Flickr 等，设置一些初始条件，验证一下线性阈值模型、独立级联模型或自选模型。

第6章

社会化媒体计算应用

本章学习目标

- 了解社会化媒体文本挖掘的情感分析
- 了解融合社会化媒体的金融预测分析
- 了解社会网络中的个性化推荐应用

6.1 基于社会化媒体文本挖掘的情感分析

6.1.1 情感分析研究概述

情感分析是指利用计算机挖掘、提取出互联网信息的褒贬态度和意见。在社会化媒体计算中，情感分析作为一种重要的分析手段，已经被广泛地应用到商务智能、舆情监控等领域中。国内外对于中英文文本的情感分析研究已经屡见不鲜。与此同时，国外对英文文本的情感分析研究主要分为“词典”、“句子”、“篇章”、“海量数据集”这四个级别。国内对中文文本的情感分析主要集中在“词语”、“句子”、“篇章”三个级别。总之，国内外众多学者对文本情感分析的研究不断深入，进一步为情感分析在社会化媒体计算中的应用做好了坚实的理论基础。

而近几年，随着微博在中国的迅速普及和用户数量的剧增，中文微博的情感分析价值也逐步凸显出来。因此，本书围绕社会化典型媒体——微博进行情感分析，为了让情感分析更具有针对性，使本书的情感分析研究更具应用扩展性，本书进一步指定围绕金融领域。以研究过程为线索，系统地阐述了基于社会化媒体的情感分析研究方法。

从总体上来说,在对特定领域(本书以金融领域为例)进行情感分析时,往往需要将研究过程进一步细分为对领域相关微博的提取和对领域相关微博情感分析两个主要步骤。第一个步骤中,对金融领域相关微博的提取等同于将微博划分为金融相关、金融不相关的二分类问题。第二个步骤中,对微博情感分析的处理同样也可以看做将金融领域微博进一步划分为正向情感金融微博、负向情感金融微博、中性情感金融微博的三分类问题。因此,如何准确有效地进行分类模型的设计是情感分析的核心。但是,在利用分类模型进行处理之前,需要对微博文本进行一定的预处理,提取重要参数指标,进而才能满足分类模型的要求。在完成情感分类之后,仍需要选取有效的指标对分类结果进行准确性相关评价。因此本节从研究过程出发,从情感分析文本预处理、情感倾向分类模型、情感分类评价指标三个模块重点介绍了情感分析中需要了解与掌握的基础研究方法。

6.1.2 情感分析文本预处理

1. 原始数据的收集

原始数据的收集是进行情感分析的基础,后续的研究都需要基于对海量微博数据准确、及时地爬取收集。微博的收集可以通过多种渠道获得,常用的两种方法可以归结为:基于专业采集软件与基于应用程序编程接口(Application Programming Interface,API)的自主编程。

在网站数据采集软件中,国内已开发了大量较为成熟的爬取工具,利用率较高的是火车头采集器 LocoySpider。数据采集软件在使用时操作简单易懂,但对数据收集结果的自动处理上仍存在软件本身的一定限制。因此,可以基于新浪微博、腾讯微博提供的 API 进行自主编程,将数据爬取结果与后续的文本处理无缝连接起来。

在数据的存储方面,也可以采用多种存储形式,如将每条微博存储为.txt 文件,或存储到数据库中等。因为考虑到微博的信息具有一定的特征,往往较为通用地记录了微博用户的昵称、粉丝数、微博内容、转发数、评论数、发布时间、用户标签等特征。因此书中采用了将微博数据存储到关系型数据库如 Microsoft SQL Server、MySQL 等之中。为下一步的微博文本处理打好相应的数据基础。

2. 微博关键词的提取

微博关键词的提取,不同于传统的文本关键词提取方式。微博的文本内容被限制在 140 字以内,并伴随有多种网络流行用语及精简的词语用法。因此,对于微博关键词的提取,主要用到了基于 TF-IDF 算法的改进方法。此方法在传统的 TF-IDF 算法基础上,增加了对词组字长的考量,使得经改进的 TF-IDF 算法得出的微博关键词更加的准确。

(1) 传统的 TF-IDF 算法简介

TF-IDF 是用来评估一个词对于一个文件集或者语料库中其所在文件的重要性。TF 词频表示某个词条在其所在文档中出现的次数。IDF 逆文档频率表示如果包含某个词条的文档越少,那么这个词条就具有很好的区分能力。词条的重要程度与其在某文档中出现的次数成正比,与其在文件集(或语料库)中出现的频率成反比。

对于某一文件中的词语 t_i 来说,其在此文件中的重要性可以表示为式(6-1)。

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{6-1}$$

其中 $n_{i,j}$ 表示这个词在文件 d_j 中出现的次数,$\sum_k n_{k,j}$ 表示此词在文件集中出现的总的次数。

对于某一文件中的词语 t_i 来说,其在文件集中的类别区分能力可以表示为式(6-2)。

$$IDF_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{6-2}$$

其中 $|D|$ 表示文件集中文件的总数,$|\{j:t_i\in d_j\}|$ 表示包含该词的文件数目。

对于某一文件中的词语 t_i 来说,其在整个文件集中的重要性可以表示为式(6-3)。

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i \tag{6-3}$$

对于特定词语 t_i 来说,其在某一文件中出现频率越高,在整个文档集中出现越低,其 TF-IDF 值越高,即在整个文件集中的重要程度越高。

(2) 针对性的 TF-IDF 算法改进

传统的 TF-IDF 算法以词频的统计为主体,忽略了特征词的位置、特征词的长度以及文件的来源是否一致等一系列的问题。从 20 世纪末开始,国内外研究者就不断地寻找方法来改进 TF-IDF 算法。改进的算法主要分为两类,一类是针对不同类别之间文档的量级不同而产生的权重计算问题。

有的学者提出了使用分类短语 CTD(Categorical Term Descriptor)的方法来改进 TF-IDF 算法,来达到修正 TF-IDF 算法在处理不同量级的文档集时的权重计算的影响。具体表示为式(6-4)和式(6-5)。

$$CTD(t_k,c_i)=TF(t_k,c_i)\times IDF(t_k,c_i)\times ICF(t_k) \tag{6-4}$$

$$ICF(t_k)=\log\left(\frac{|C|}{CF(t_k)}\right),\quad IDF(t_k,c_i)=\log\left(\frac{|D(c_j)|}{DF(t_k,c_i)}\right) \tag{6-5}$$

其中 $TF(t_k,c_i)$表示特征项 t_k 在类 c_i 中出现的次数;$D(c_j)$表示类别 c_j 中的文档

数；$DF(t_k, c_i)$指类别 c_i 中出现特征项 t_k 的文档数；C 代表类别数。

有的学者认为传统的 TF-IDF 算法没有考虑到特征项在类内和类之间的分布情况，针对这种情况提出了一种结合信息熵的改进方法。该方法通过信息分布熵来调整 TF-IDF 的特征圈子，避免了对分类没有太大贡献的特征项赋予较大的权值，从而提高文本分类的精度和召回率。其具体表示如式(6-6)所示。

$$W_{ik}(d) = \frac{tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_{i=1}^{n} (tf_{ik}(d))^2 \times \left[\log\left(\frac{N}{n_k} + 0.01\right)\right]^2}} \times \alpha(H_{ac}) \times H_{ic} \tag{6-6}$$

其中 $\alpha(H_{ac})$表示特征项在类间的分布情况，H_{ic} 表示特征项在类内的分布情况。当一个特征项在类间分布越均匀，此类间分布熵越大，$\alpha(H_{ac})$越小，对分类贡献就越小。当某一特征项在类内分布越均匀，类内分布熵H_{ic}越大，对分类贡献就越大。

另一类，是通过引入新的因子来改进 TF-IDF 算法。

有的学者引入对特征词词长和位置的考虑来改进 TF-IDF 算法，并用《半导体光电》杂志的真实数据为例，验证了改进后的算法的确能够提高特征词抽取的效率和准确性，结果如式(6-7)所示。

$$w_t = \frac{(fre_{t1} \times \lambda_{t1} + fre_{t2} \times \lambda_{t2} + fre_{t3} \times \lambda_{t3} + L) \times \log\left(\frac{N}{n_t} + 0.01\right)}{\sqrt{\sum_{t=1}^{m} \left[(fre_{t1} \times \lambda_{t1} + fre_{t2} \times \lambda_{t2} + fre_{t3} \times \lambda_{t3} + L) \times \log\left(\frac{N}{n_t} + 0.01\right)\right]^2}} \tag{6-7}$$

其中 w_t 为考虑词长和位置的特征词所得权重；fre_{t1}、fre_{t2}、fre_{t3}表示特征词 t 在文档标题、关键词、摘要部分出现的频数，λ_{t1}、λ_{t2}、λ_{t3} 表示特征词 t 出现在上述三个位置的权重系数，L 为词长权重；n_t 表示特征词 t 出现文档的频数；N 为文档集中的文档数目；m 为特征词数目。

因为研究对象为微博数据，而微博一般为 1～3 句话(140 字以内)构成，所以引入词长来改进 TF-IDF 算法比较适宜。如式(6-8)所示。

$$\begin{aligned} TFIDF_{i,j} &= TF_{i,j} \times IDF_i \times L \\ &= \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{j : t_i \in d_j\}|} \times L \end{aligned} \tag{6-8}$$

其中，$n_{i,j}$ 表示这个词在文件 d_j 中出现的次数，$\sum_k n_{k,j}$ 表示此词在文件集中出现的总的次数；$|D|$表示文件集中文件的总数，$|\{j : t_i \in d_j\}|$ 表示包含该词的文件数目；L 为词长权重。

3. 基于 ICTCLAS 的文本分词处理

文本的情感分析往往是在词语级别基础上进行的，因此文本的分词处理，同时标注词语词性是进行情感分析之前十分重要的预处理内容。分词的结果将用于基于改进的 TF-IDF 方法的关键词抽取，以及基于多维度词典的微博情感特征模块识别的分析中去。

目前，国内的汉语分词系统有很多，常见的中文分词开源项目有 SCWS、ICTCLAS、HTTPCWS、庖丁解牛分词和 CC-CEDICT 这五种。HTTPCWS 是基于 HTTP 协议的中文分词系统，目前只能在 Linux 下运行。CC-CEDICT 是以汉语拼音为附中的英汉词典为基础进行中文分词的系统。而 ICTCLAS 是我国最早的分词系统，来源于中国科学院计算技术研究所，并由张华平博士进行升级、完善。其主要功能包括命名识别、词性标注、中文分词、新词识别，并且支持用户自定义词典，支持 UTF-8、BIG-5 以及 GBK 等不同的编码。目前 ICTCLAS 系统已经升级到了 ICTCLAS 3.0。ICTCLAS 3.0 分词速度单机 996KB/s，分词精度 98.45%，API 不超过 200KB，各种词典数据压缩后不到 3MB，是当前世界上最好的汉语词法分析器。在国家 973 计划项目评测中，ICTCLAS 的针对简体中文的分词效果要明显优于其他系统。ICTCLAS 的分词原理结构图如图 6.1 所示。

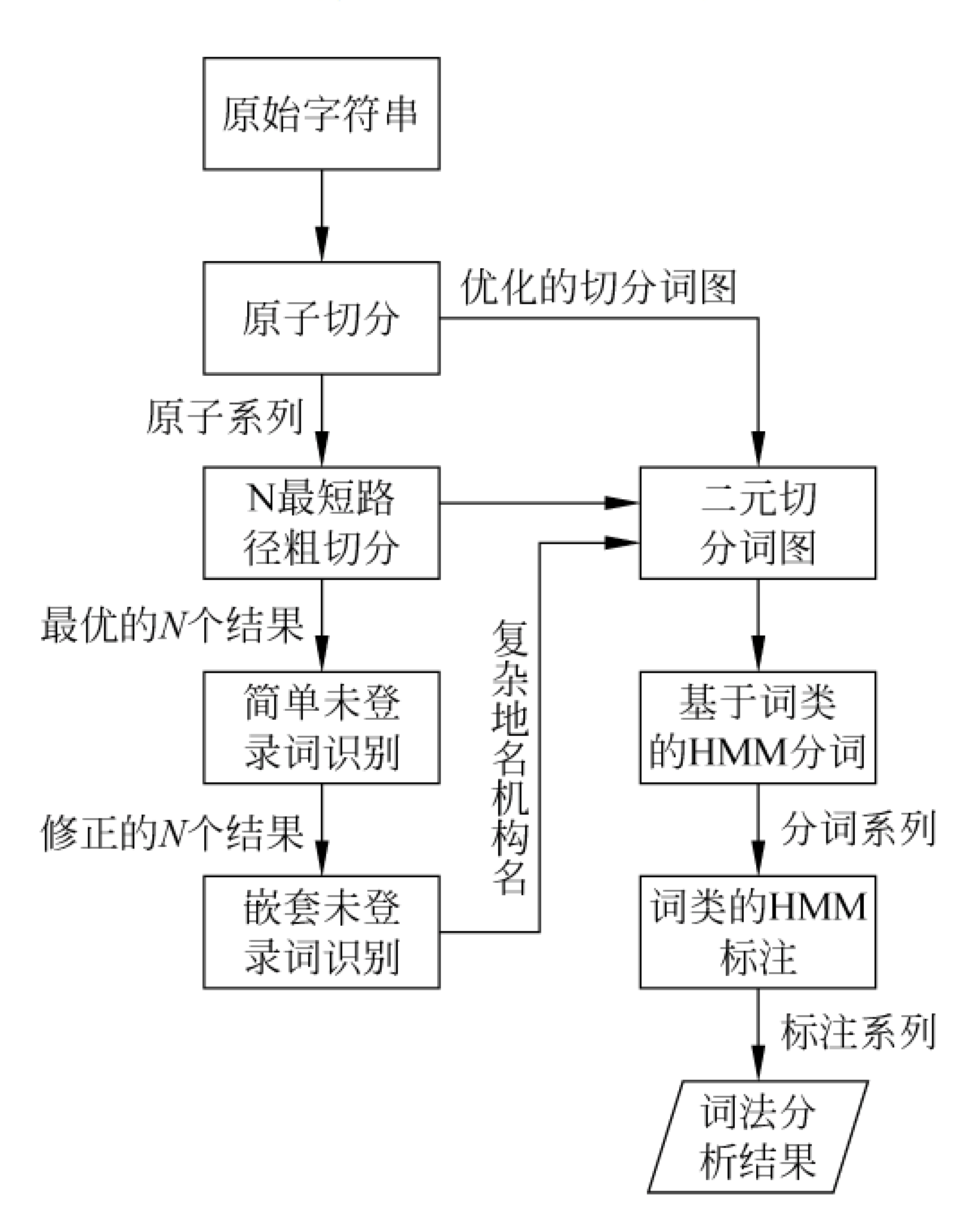


图 6.1　ICTCLAS 的框架结构

（资料来源：张华平，语言浅层分析与句子级新信息检测研究，中国科学院计算技术研究所）

对于微博内容的分词可以采用NLPIR汉语分词系统(ICTCLAS)2013 Java版。此版本新添加了针对微博的分词功能,能够有效地识别微博中的特定用词,如"给力"等。对微博文本进行分词的同时,还能对每个分好的词进行词性的标注(采用计算机所一级标注)。

ICTCLAS 2013版的计算所的词性标注集说明如表6.1所示。

表6.1 ICTCLAS 2013词性标注集表

词性	代码	词性	代码	词性	代码	词性	代码
名词	n	时间	t	处所词	s	方位词	f
动词	v	形容词	a	区别词	b	状态词	z
代词	r	数词	m	量词	q	副词	d
介词	p	连词	c	助词	u	叹词	e
语气词	y	拟声词	o	前缀	h	后缀	k
字符串	x	标点符号	w				

6.1.3 微博情感倾向分类模型

微博的情感分类算法大体可以分为两类:基于规则的方法与基于机器学习的方法。其中,基于规则的方法是指,以指定的情感词典为基础,将微博的文本内容与情感词典匹配出的特征极性进行加和求总,最终得到的结果为整条微博的极性,因此判断微博的情感倾向。而基于机器学习的方法,则是基于微博文本提取出的特征值作为机器学习的输入属性,对分类模型进行训练,并将训练好的模型作为主要分类手段的方法。

1. 基于规则的方法

基于规则的方法往往建立在多维度词典的基础之上,根据文本预处理得到的微博分词结果与情感词典中的极性特征词进行匹配,每一个可以与情感词典相匹配的词都被赋予一定的极性值,最终把所有极性值按一定的权重相加,得到每一条微博的情感极性值。因此,在基于规则的方法中,对情感词典的选取非常重要,情感词典的建立可以利用已有的较为全面的词典,也可以根据需要自建词典,本书主要选取了以下三个典型的情感词典为例进行详细介绍。

(1) 基于表情词典的特征选择

在微博语言中,各式各样的表情符号能够很直观地表现出微博发布者的情绪特点。针对微博的情感分析,我们收集了腾讯微博的情感符号,并制成了表情符号字典。图片格式的表情符号被存储成文本格式,例如☺表示为"/微笑",分为正向表情字典和负向表情两类。正向表情字典共收录23个表情符号,负向表情字典共收录27个表情符号。构建的表情字典如表6.2所示。

表 6.2　腾讯微博表情符号字典

词典	内　　容
bqz.txt(正向)	得瑟、给力、微笑、色、得意、调皮、龇牙、强、胜利、OK、跳跳、转圈、酷、可爱、憨笑、奋斗、鼓掌、亲亲、太阳、捶地大笑、左太极、右太极、飞吻
bqf.txt(负向)	压力山大、撇嘴、流泪、闭嘴、大哭、尴尬、发怒、难过、吐、白眼、弱、没心情、伤不起、怄火、惊恐、咒骂、折磨、衰、骷髅、敲打、鄙视、菜刀、炸弹、刀、石化、叹气、发抖

用表情字典做特征项的选择，关键要素是分析微博的存储结构。正常的微博经 ICTLAS 分词后，按条存储在 txt 文档之中。但由于分词系统的不完备性，一些表情符号的文本会被错分成几个词。例如："大哭"经分词后在 txt 文档中储存为"大哭"三个分开的部分。为了准确提取此类表情分行，首先将分词后的微博储存在 string[]中，用空格来区分词组。其次，寻找"/"，寻找到"/"后，再分别往后遍历1、2、3、4 位，与"/"重新组合成词组，再放入正向表情和负向表情字典中进行匹配，并计算其词组在字典中出现的次数。

(2) 基于 Hownet 情感词典的特征选择

以 Hownet 情感字典为基础进行特征提取，主要思想是首先将分词后的微博储存在 string[]中，用空格来区分词组。若是这两个字典中有这个词，则向这个词的前后各寻找两个词，并将其带入程度词字典和否定词前缀字典进行匹配，如图 6.2 所示。

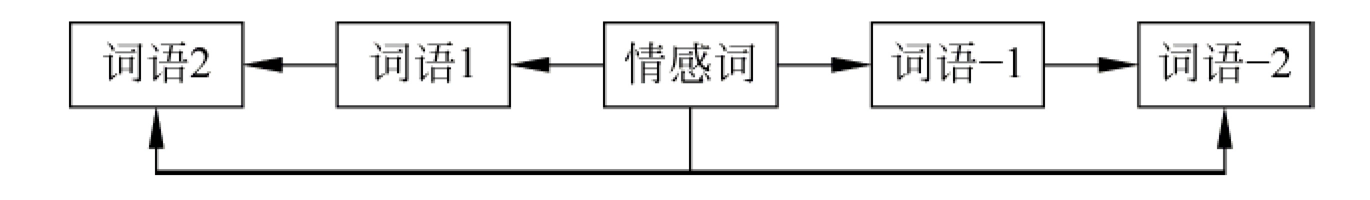


图 6.2　基于 Hownet 词典判断短语极性示意图

Hownet 情感字典收录了正向的情感词、负向情感词、否定词前缀、六个级别的程度词。正向的情感词包括"蔼然、安如磐石、昂然"之类的词，负向的情感词包括"哀鸿遍野、瘆得慌、暗无天日"之类的词，否定词前缀包括"不能、不应该、不"之类的词。而六个级别的程度词，由完全肯定(否定)到相对肯定(否定)被分成了六个级别词典。具体如表 6.3 所示。

考虑到负向情感词对人影响往往大于正向情感词，所以将正向情感词定基为1，负向情感词定基为－1.5。若是该词的前后两位有程度词，则根据不同的程度，给予不同的系数权重。若是该情感词的前两位有否定前缀字典中的词，则乘以系数－1。

表 6.3　Hownet 情感词典

词典	内　　容
privative. txt（否定前缀词典）	不能、不应该、不、不得、不让、不然、不就
positive. txt（正向情感词典）	蔼、蔼蔼、蔼然、蔼然可亲、蔼如、艾、碍事、安、安分、安好、安静、安康安澜、安乐、安宁、安全、安然、安然无事……
negative. txt（负向情感词典）	瘆得慌、哀鸿遍野、矮、碍难、碍眼、爱搭不理、爱理不理、暗、暗淡、暗地、暗地里、暗黑、暗里、暗昧、暗无天日、暗下、暗中……
degree1. txt（程度词典）	百分之百、倍加、备至、不得了、不堪、不可开交、不亦乐乎、不折不扣、彻头彻尾、充分、到头、非常、极、极度、极端、极其、极为
degree2. txt（程度词典）	不过、不少、不胜、惨、沉、沉沉、出奇、大为、多、多多、多加、多么、分外、格外、够瞧的、好、好不、何等、颇、颇为、甚、实在……
degree3. txt（程度词典）	大不了、多、更、更加、更进一步、更为、还、还要、较、较比、较为、进一步、那般、那么、那样、强、如斯、愈发、愈加、愈来愈……
degree4. txt（程度词典）	不为过、超、超额、出头、多、浮、过、过度、过分、过火、过劲、过了头、过猛、过热、过甚、过头、过于、苦、老、偏、强、溢、忒……
degree5. txt（程度词典）	点点滴滴、多多少少、怪、好生、还、或多或少、略、略加、略略、略微、略为、蛮、稍微、稍为、稍许、挺、未免、相当、些、些微……
degree6. txt（程度词典）	半点、不大、不丁点儿、不甚、不怎么、聊、没怎么、轻度、弱、丝毫、微、相对

（3）基于网络用语词典的特征选择

互联网是一个口语语言的集中地，不仅如此，互联网还形成了自己独特的语言特点。例如用“稀饭”表示喜欢，“给力”、“8 错”表示赞扬，high 表示特别高兴，“悲催”、“悲剧”、“给跪了”表示强烈的负面情感，“呵呵”表示不认可、无奈、轻视的情绪。这些词语具有很强的情感表现力，而 Hownet 之类通用的词典并没有对其进行收录，所以文章在此基础之上又结合网络用语，提取出基于网络用语词典的文本特征。

下面人工收录了近些年网络上流行的用语，并提取出具有明显情感倾向的词语构成网络用语正向和负向情感字典，作为补充词典。构建的网络用语情感字典如表 6.4 所示。wlyyz. txt 为正向的网络用语词典，wlyyf. txt 为负向的网络用语词典。

表 6.4　基于网络用语的情感词典

词典	内　　容
wlyyz. txt(正向)	顶、狂顶、流口水、happy、high、小强、养眼、大虾、8 错、稀饭、果酱、走召弓虽、嘻嘻、gx、NB、弓虽、牛 x、有料、CM、给力、牛 B、高富帅、白富美
wlyyf. txt(负向)	晕、靠、拍砖、衰、恐龙、废柴、屌丝、BT、FT、SL、MD、TMD、TNND、JJWW、SJB、PMP、MPJ、抓狂、包子、蛋白质、55555、BC、JR、JS、垃圾、泪、呵呵、kao、damn、倒、寒、SIGH、DBC、puke、SB、BS

网络用语情感词典在情感分析中的使用方法类似于 Hownet 词典的使用方法。在用程序实现过程中，首先将分词后的微博储存在 string[]中，用空格来区分词组。将存储微博的数组代入正向情感和负向情感字典中进行遍历。若是这两个字典中有这个词，则向这个词的前后各寻找两个词，并将其代入程度词字典和否定词前缀字典进行匹配。为了使得分词时不把这些网络用词给拆分开，故在分词程序中添加用户词典，导入网络用词。在进行情感倾向特性选择时，将网络用语词典并入 Hownet 词典中进行情感倾向统计。

总而言之，微博情感倾向的特征选择都是基于网络用语词典、Hownet 情感词典、表情词典三维度字典之上完成的，分析过程为对同一篇微博文本依次进行基于三个字典的特征统计。并以统计的结果作为每一条微博机器学习情感倾向分类的输入值。

2. 基于机器学习的算法

与基于规则的算法相比，基于机器学习的算法往往能够表现出更好的分类结果，机器学习算法中较为典型的算法理论有支持向量机(Support Vector Machine，SVM)、BP 神经网络算法、朴素贝叶斯算法、决策树(Decision Tree)算法等。基于以往的研究成果，SVM 算法一般能够表现出更好的分类结果，因此本书中以 SVM 算法为例对机器学习的方法进行阐述。

(1) 支持向量机

支持向量机由 Vapnik 等在 1995 年提出，SVM 在解决小样本、非线性识别中具有巨大的优势。支持向量机的最根本的原理就是将低维空间中的点映射到高维空间中，使之成为线性可分的。再利用线性划分原理进行分类边界的判断。

对于线性可分的二分类问题，可以选择直接用线性的支持向量机分类机。其公式表示如下。

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{j=1}^{l}\alpha_j$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i$$

$$\alpha_i \geqslant 0 \quad i = 1,\cdots,l$$

据此可以得到式(6-9)。

$$W^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i, \quad b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i (x_i \cdot x_j) \tag{6-9}$$

对于线性不可分问题，可使用线性软间隔分类机、非线性硬间隔分类机、C-支持向量分类机以及 V-支持向量分类机等。最常用的是 C-支持向量机，其分类问题可以表示为

$$\text{映射：} T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$$

$$\text{且 } x_i = \phi(x_i)$$

$$\text{分类面：} (w \cdot \tilde{x}) + \tilde{b} = 0$$

$$\min_{w,b,\xi} \frac{1}{2} w^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s. t. } \; y_i((w \cdot x_i) + b) \geqslant 1 - \xi_i, i = 1, \cdots, l$$

$$\xi_i \geqslant 0, \quad i = 1, \cdots, l$$

与对偶问题

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j$$

$$\text{s. t. } \; \sum_{i=1}^{l} y_i \alpha_i$$

$$0 \leqslant \alpha_i \leqslant C, i = 1, \cdots, l$$

据此可得式(6-10)。

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_j)$$

$$f(x) = \operatorname{sgn}\left(\sum \alpha_i^* y_i K(x, x_i) + b^* \right) \tag{6-10}$$

为了使 SVM 简单易用，台湾大学林智仁教授开发了 libsvm 软件包。并提供了源码，方便修改和调用。因此，可以在 MATLAB 的平台上调用 libsvm 的软件包，对微博的情感倾向进行分类实现。

(2) 支持向量机在微博情感分析中应用

在对金融相关领域进行情感分类过程中，由上文可知，该过程可以分为提取金融相关微博与金融微博情感分类两个步骤。以第一步提取金融相关微博为例，整个过程如图 6.3 所示。

在进行分类模型输入特征选取上，基于微博文本特点，挑选了微博的标点数(Bdcount)、标点类别(Bdlb)、文本长度(Allcount)、在自建金融词典中出现的次数(Jrcount)、在搜狗金融词典中出现的次数(Qfcount)作为输入维，以微博的类别(金融相关微博为 1，非金融相关微博为 0)作为输出维。输入向量如式(6-11)所示，其中 $x_{n,m}$ 表示第 n 条微博的关于金融领域相关特征的第 m 个特征项。

微博.txt 文本文件 —提取特征 输入维→

	bdco	*bdlb*	*allco*	*jrco*	*qfco*
第1条	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$
第2条	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$
⋮	⋮	⋮	⋮	⋮	⋮
第n条	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	$x_{n,5}$

$$x_{i,j}=\frac{x_{i,j}-\min(x_{i,j})}{\max(x_{i,j})-\min(x_{i,j})}$$

数据标准化 St.$v_{i,j}\in[0,1]$ →

	bdco	*bdlb*	*allco*	*jrco*	*qfco*
第1条	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$
第2条	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$
⋮	⋮	⋮	⋮	⋮	⋮
第n条	$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	$y_{n,4}$	$y_{n,5}$

SVM 分类算法 x_i取0,1 →

第1条	x_1
第2条	x_2
⋮	⋮
第n条	x_n

统计 → 1金融相关=()条 0金融不相关=()条

图 6.3 金融领域微博提取流程图

$$X_{input2}=\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} & \cdots & x_{n,m} \end{bmatrix} \tag{6-11}$$

对于样本数据来说，每个属性的量级都有可能不一样。若是不对数据做标准化处理，量级大的数据影响会被放大，而量级小的数据影响将会被缩小。为了规避此种情况，首先对数据集进行归一化处理，将数据的值规范化到[0，1]之间。

对于属性 $x_{i,j}$来说，其标准化到[0，1]区间，如式(6-12)所示。

$$x'_{i,j}=\frac{x_{i,j}-\min(x_{i,j})}{\max(x_{i,j})-\min(x_{i,j})} \tag{6-12}$$

其中，$x_{i,j}$为数据集属性 i 的第 j 个值。$\min(x_{i,j})$为属性 i 的最小值，$\max(x_{i,j})$为属性 i 的最大值。分类模型的输出如式(6-13)所示。

$$Y_{output1}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{6-13}$$

其中，y_n为分类标签，若为金融相关微博则结果为 1，反之为－1。

将整个已标注的微博数据集(Dataset)分为训练集(Training Set)和测试集(Test Set)。使用 SVM 进行分类试验，SVM 的参数：－s 0，－t 2。根据测试集的结果进一步完成下一模块对分类结果评价的分析。

而对金融微博的情感分类与第一步方法相似，但是因为金融微博情感分类是一种包含正向情感、负向情感和中性情感的三分类问题，因此情感分类最终分为三类：正向、负向和中性。故选择使用一对一(One Versus One)的方法进行分类训练。具体说来，就是将此三分类转化为三个二分类问题，即"正向—负向"，"正向—中性"，"负向—中性"三个分类。将此三个类别分别进行训练得到三个训练模型，再将测试集代入这三个模型之中进行测试，最后再采用投票的方式，决定测试集中的每条微博到底属于哪个类别。

投票的过程可以表示为，

A＝B＝C＝0，A 代表正向类别的值，B 代表负向类别的值，C 代表中性类别的值；

(正向，负向)—分类器：如果属于正向则 A++，如果属于负向，则 B++；

(正向，中性)—分类器：如果属于正向则 A++，如果属于中性，则 C++；

(负向，中性)—分类器：如果属于负向则 B++，如果属于中性，则 C++；

每条微博类别的最终归属为 A，B，C 中最大的值。

6.1.4 情感分类评价指标

对于分类问题的评价指标，一般包括正确率、召回率和 F 指数这三个。

1. 正确率

正确率即查准率，此指标表示实验对目标判断准确的能力，正确率越高，则误判的概率越小，

$$P_1 = \frac{\text{正确判断为金融相关微博数}}{\text{判断为金融相关的微博数}}$$

$$P_2 = \frac{\text{正确判断为正向情感的微博数}}{\text{判断为正向情感的微博数}}$$

$$P_3 = \frac{\text{正确判断为负向情感的微博数}}{\text{判断为负向情感的微博数}}$$

$$P_4 = \frac{\text{正确判断为中性情感的微博数}}{\text{判断为中性情感的微博数}}$$

对于提取金融相关微博的准确率用 P_1来表示，即正确判断为金融相关微博的条数与判断为金融相关微博条数的比值。对于微博情感极性的判断准确率，分为正向情感和负向情感和中性情感三类。正向情感的判断正确率用 P_2来表示，即正向情

感判断正确的微博数与判断为正向情感微博数的比值。负向情感的判断正确率用 P_3 来表示，即负向情感判断正确的微博数与判断为负向情感微博数的比值。中性情感的判断准确率用 P_4 表示，即中性情感判断正确的微博数与判断为中性情感的微博数的比值。

2. 召回率

召回率即查全率，此指标反映了实验发现目标的能力。召回率越高，则漏判的越少。本实验中相关公式表示如下：

$$R_1=\frac{\text{正确判断为金融相关微博数}}{\text{实际为金融相关的微博数}}$$

$$R_2=\frac{\text{正确判断为正向情感微博数}}{\text{实际为正向情感的微博数}}$$

$$R_3=\frac{\text{正确判断为负向情感微博数}}{\text{实际为负向情感的微博数}}$$

$$R_4=\frac{\text{正确判断为中性情感微博数}}{\text{实际为中性情感的微博数}}$$

对于提取金融相关微博的召回率用 R_1 来表示，即为正确判断为金融相关微博的数量与实际为金融相关微博数的比值。对于微博情感极性的召回率，分为正向情感、负向情感和中性情感三类类。正向情感微博的召回率用来 R_2 表示，即正确判断为正向情感的微博数与实际为正向情感的微博数的比值。负向情感微博的召回率用 R_3 来表示，即正确判断为负向情感的微博数与实际为负向情感的微博数的比值。中性情感微博的召回率用 R_4 表示，即正确判断为中性情感的微博数与实际为中性情感的微博数的比值。

3. *F* 指数

F 指数是用来综合衡量准确率和召回率的指标。定义如式(6-14)所示：

$$F=\frac{(\beta^2+1.0)\times P\times R}{\beta^2\times P+R} \tag{6-14}$$

β 表示准确率和召回率的相对权重。当 β 等于 1 时，两者同样重要；当 β 大于 1 时，召回率更加重要；当 β 小于 1 时，正确率的比重更高。在 β 取相同值时，F 指数的值越高，则说明此种算法越好。就研究对象而言，正确率更显重要，因此 β 取小于 1 的值，定为 0.5。

6.2　基于流形学习的社会化媒体金融复合数据的预测

6.2.1　金融预测研究概述

中国的股票市场经过了多年的发展和演化，不断地自我完善，逐步成熟，但是

依然存在着一些问题。股票市场的波动是一把双刃剑,可能为民众带来丰厚的回报,也有可能造成巨大的社会问题。而在已有的研究中,学者们通过很多不同的方法来探究股票市场的发展规律,但是由于股票市场本身的复杂性,股票时间序列的非平稳性、长尾性以及影响因素众多等特点,准确预测股票价格的走势是一项艰巨的任务。

近年来,网络舆情对股市的影响也逐步凸显出来,投资者对于有公众新闻的股票容易反应过激。同时,雅虎等网站上的股票新闻能够显著影响股票投资的收益。种种研究同时表明,网络舆情的信息量也能对股价的预测产生一定影响。此外,Google 相应关键词的搜索能够对道琼斯大盘指数进行短期的有效预测,基于 2004—2011 年的数据利用模拟实验实现了超过 300%的投资回报率。与此同时,就中国资本市场而言,上交所规定了紧急停牌重点舆情监测媒体 15 家:上海证券报、中国证券报、证券时报、第一财经日报、21 世纪经济报道、经济观察报、证券日报、华夏时报、每日经济新闻、中国经营报、财经、证券市场周刊、新世纪周刊、和讯、新浪财经等,实际意义上证明了网络舆情对于股票市场的影响能力。因此,本章着重对基于网络舆情的股市价格预测方法做了较为全面的阐述。

在社会化媒体计算的视角下,从社会化媒体信息角度出发,利用互联网文本信息、数据信息等多渠道获得的数据,提供一个全新视角的股市预测分析方法。

6.2.2 原始数据获取及量化处理

1. 原始数据的获取

基于互联网的舆情信息来源较为广泛,综合国内外研究的成果,本书考虑可以从获取股票价格时间序列数据、网络搜索词热度、股票新闻舆情信息数据三种数据源作为原始数据获取的途径。其中股票价格时间序列是逐日股票交易产生的基础数据,分别为开盘价、最高价、最低价、收盘价、成交量、成交额,数据可以由金融数据服务商提供(如 Wind 资讯);网络搜索词热度可以根据百度指数对于定义搜索词的热度时间序列。股票新闻舆情数据是由网络搜集而来的股票新闻数据,将舆情信息统计后获取信息量和情感强度两个基础时间序列得来。

(1) 股票价格时间序列数据获取

对于股票价格时间序列数据而言,主要的股票技术指标基础数据有六个,分别是开盘价,最高价,最低价,收盘价,成交量(手),成交额(百万元)。源数据可以由多种渠道搜集,本章数据主要由资讯金融终端获得,在此基础上可以根据需求,对不同的股票市场和板块市场进行分析。本节参考中国人民大学信息学院经济信息管理系林航同学硕士论文(林航,2013),针对中国整个股市及银行板块市场为例展开分析,收集两个典型综合指数(沪深 300 综合指数、银行板块指数)以及四支银行

板块的个股(民生银行—600016、兴业银行—600036、招商银行—601166、中信银行—601998)。其中,选取个股的依据为股份制银行中利润总额最高的四家。选取申银万国证券一级行业分类指数——银行(申万)作为行业指数数据,同样提取开盘价,最高价,最低价,收盘价,成交量(手),成交额(百万元)等数据。最终,再获取沪深300指数。

(2) 网络搜索热度数据获取

对于网络搜索词热度数据的获取,可以利用百度指数平台获取对于相应关键词的搜索热度。百度指数平台如图6.4所示。



图6.4 百度指数平台网络页面

利用关键词——沪深300、银行业、民生银行、兴业银行、招商银行、中信银行分别进行逐月检索,获取关键词搜索的逐日热度,形成相应搜索词热度时间序列,

以用于下一步处理。

（3）股票新闻舆情信息数据获取

对于股票新闻舆情信息数据的处理，可以利用各大新闻门户网站和独立财经媒体获取互联网财经类新闻。主要从三种新闻类型进行数据的抓取，一是影响大盘走势的新闻（以下简称市场新闻），二是影响行业走势的新闻（以下简称行业新闻），三是影响个股走势的新闻（以下简称个股新闻）。抓取的所需数据为网站HTML源代码中的有效文本，每个网站的代码存在极大差异，到目前为止还没有简单的方法能准确地从网页中直接抽取所需的文本信息。因此，可以采用整站抓取的策略，利用网络爬虫Heritrix。Heritrix的整体结构如图6.5所示。Heritrix是一个爬虫框架，从总体而言，其为一个平台结构，各组成部分都具备松散耦合的特点，可以自定义地重组，为基于Heritrix的二次开发提供了基础。

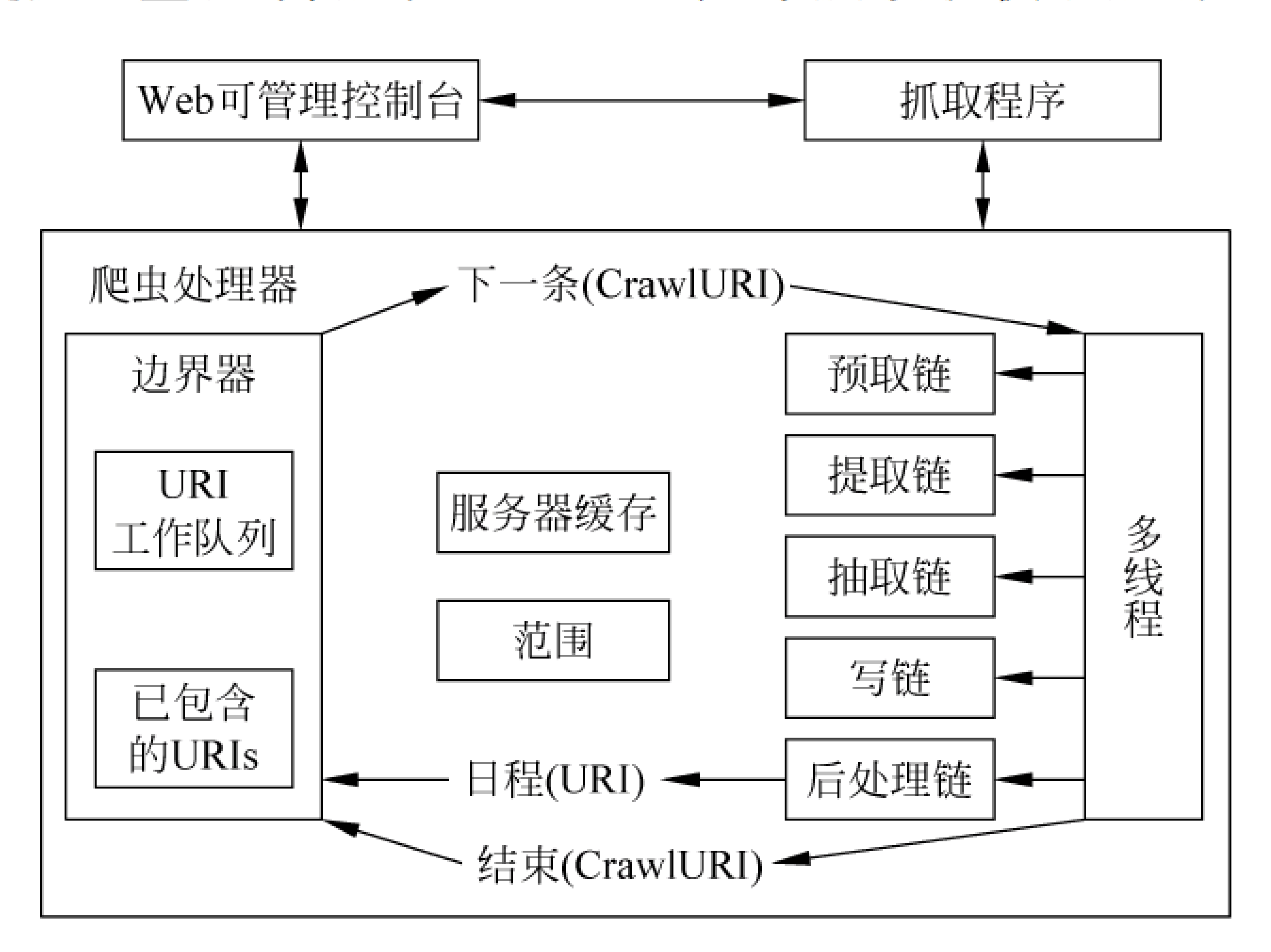


图6.5 Heritrix整体结构

（资料来源：Mohr等，Introduction to Heritrix，2004）

Heritrix工作流程如图6.6所示。每个URI都有一个独立的线程，边界控制器（Frontier）将爬过的URI标记，同时将未处理过的链接放入等待处理的Processor Chains（处理链）中采用多线程处理，Toe Thread代表处理URI的线程，最后经过一系列Processor（处理器）处理获得所需数据。

利用Heritrix对具体站点的信息采集中，这些站点可能存在众多外链，所以可能会产生采集到很多其他无关页面的数据冗余情况，这无疑会大大降低采集效率，因而针对不同的站点，需要定义相应的网址筛选规则，以确保不会采集到其他无关页面。解决这个问题的具体处理方法有两种，一是向Heritrix添加自定义的Extractor来限制解析出来的URL，二是扩展PostProcessor，对进入待处理队列的URL进行筛选处理，防止无关的链接进入队列。

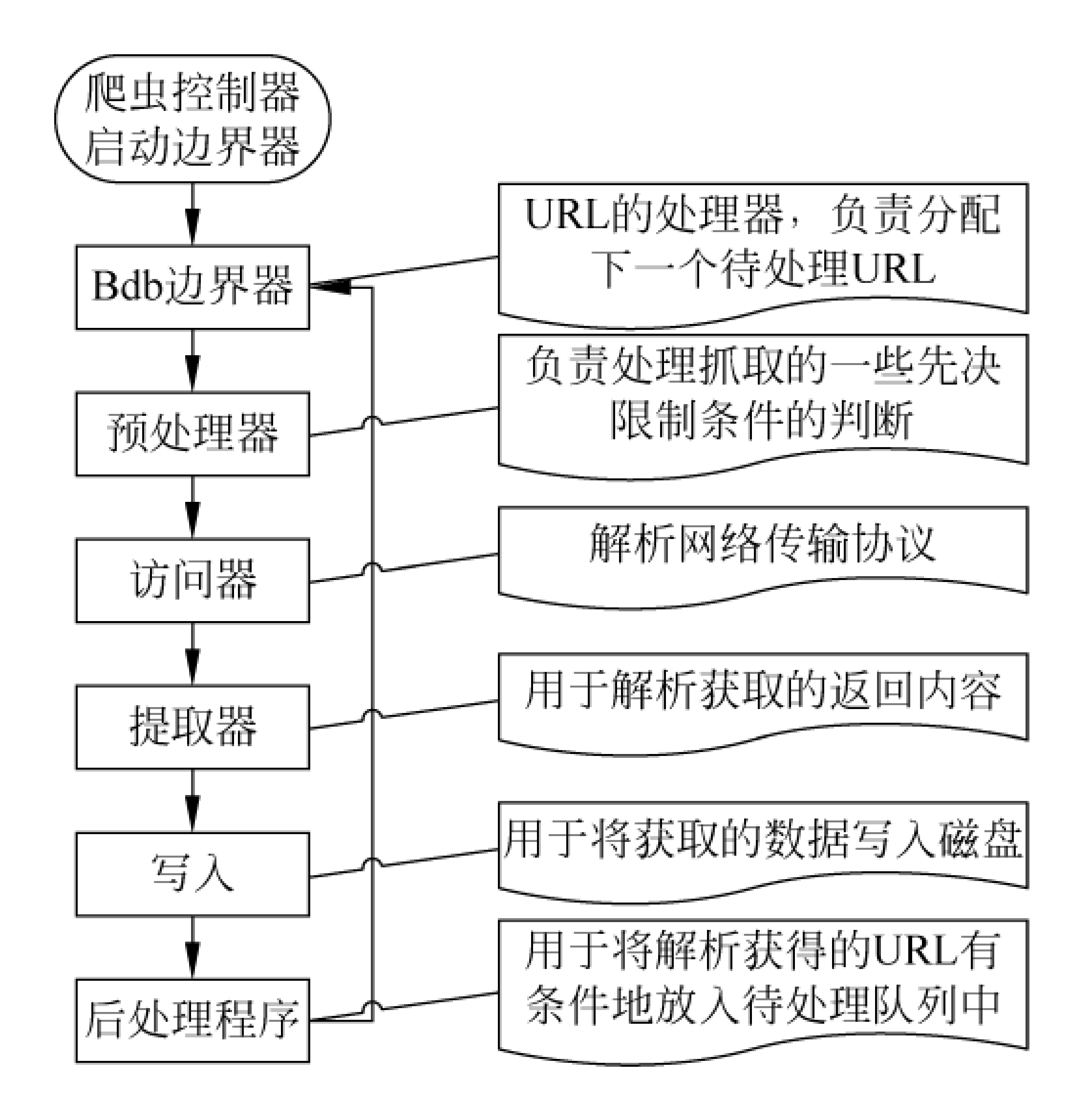


图 6.6　Heritrix 工作流程

在对 Heritrix 进行二次开发后，抓取 HTML 源代码，再利用 Java 编程语言，在 Eclipse 平台中将 HTML 语言的冗余字段删除，最终获得互联网财经类新闻文本的数据抓取，并利用 SQL Server 或其他型数库存储最终数据。

2. 原始数据的量化处理

在完成原始数据的获取后，需要分别对三类数据进行量化处理。进而完成后续数据进一步的处理分析。

(1) 股票价格时间序列数据的量化

仅仅采用单个或几个技术指标作为输入变量往往存在一定的片面性，只有通过多种技术指标的组合输入才能提高预测模型对于股票价格时间序列的预测能力。同时，考虑到预测需要抓住短期的波动规律，因此，选取表 6.5 中的技术指标作为输入指标候选。

表 6.5　技术指标详表

代码	描　　述
X_1	前一天的开盘价
X_2	前一天的最高价
X_3	前一天的最低价
X_4	前一天的收盘价
X_5	前一天的成交额(百万)
X_6	前一天的成交量(股)
X_7	BIAS1：乖离率 1

续表

代码	描　　述
X_8	BIAS2：乖离率 2
X_9	CCI：顺势指标。$TYP_i=\frac{H_i+L_i+C_i}{3}$，$CCI=\frac{TYP_i-MA(TYP,3)}{0.015*AVEDEV(TYP,3)}*100$，其中，AVEDEV 代表求平方绝对误差
X_{10}	PDI：上升方向线
X_{11}	MDI：下降方向线
X_{12}	ADX
X_{13}	K 线：RSV 的 3 日移动平均。$RSV=\frac{C_i-L_n}{H_n-L_n}*100$，其中 C_i 为第 i 日收盘价；L_n 为 n 日内的最低价；H_n 为 n 日内的最高价
X_{14}	D 线：K 值的 3 日移动平均
X_{15}	J 线：3×D－2×K
X_{16}	MACD：指数平滑异同移动平均线指标。MACD＝(DIF－DEA)/2，DIF＝EMA3－EMA6，DEA＝EMA(DIF，3)
X_{17}	PSY：心理线。PSY＝N 日内上涨天数/N＊100
X_{18}	RSI：相对强弱指标
X_{19}	SAR：停损点转向指标。SAR(n)＝SAR(n－1)＋AF[EP(N－1)－SAR(N－1)]，AF 为加速因子(或叫加速系数)，EP 为极点价(最高价或最低价)
X_{20}	ROC：变动速率。ROC＝(今天的成交均价－N 日前的成交均价)/N 日前的成交均价×100
X_{21}	BBI：多空指数。BBI＝(1 日均价＋2 日均价＋3 日均价＋4 日均价)÷4
X_{22}	LWR1：威廉指标实际上是 KD 指标的补数，100－线 K
X_{23}	LWR2：威廉指标实际上是 KD 指标的补数，100－线 D
X_{24}	DPO：区间震荡线。收盘价减收盘价的 6 日均线在 3 天前的值
X_{25}	VROC：量变动速率指标。成交量减 3 日前的成交量，再除以 3 日前的成交量，放大 1 倍，得到 VROC 值
X_{26}	DN：波幅通道下线。下限＝(收市移动平均价－波动幅度)×K
X_{27}	WR：威廉指标
X_{28}	SI：Siswing Index
X_{29}	MJR：比较当天收市价与昨天收市价的关系
X_{30}	ALF：过滤指标
X_{31}	SMI

(2) 网络搜索热度数据量化

获取网络搜索词热度、网络新闻舆情情感值强度和网络新闻舆情数量三个基本指标，构造如表 6.6 几个输入指标作为新的预测变量输入。

表 6.6　网络舆情指标详表

代码	描　　述
X_{32}	网络搜索词热度
X_{33}	网络搜索词热度每日变化倾向。计算方法为计算相邻两个时间点之间的网络搜索词热度斜率
X_{34}	网络搜索词热度威廉指标。3 日内网络搜索词热度最高值与当日值之间的差，除以 3 日内最高值与最低值的差
X_{35}	网络搜索词热度变动规则。利用 4 日内网络搜索词热度变动和 3 日内股票价格时间序列变动设立股票方向变动规则，利用规则获取次日股票移动方向推测值
X_{36}	网络新闻舆情情感值
X_{37}	网络新闻舆情情感值每日变化倾向。计算方法为计算相邻两个时间点之间的网络新闻舆情情感值斜率
X_{38}	网络新闻舆情情感值威廉指标。3 日内网络搜索词热度最高值与当日值之间的差，除以 3 日内最高值与最低值的差
X_{39}	网络新闻舆情数量
X_{40}	网络新闻舆情数量每日变化倾向。计算方法为计算相邻两个时间点之间的网络新闻舆情数量斜率
X_{41}	网络新闻舆情数量威廉指标。3 日内网络新闻舆情数量最高值与当日值之间的差，除以 3 日内最高值与最低值的差

（3）股票新闻舆情信息数据的量化

互联网财经类新闻文本处理主要分为两个部分，第一个部分对之前处理过的有效文本信息进行分词，第二个部分则是对分词后的内容计算情感值，用于指标的输入。

由于获取到的文本信息是一个连续的文本，如果想对其进行量化需要首先进行分词处理。可以采用上一节中所提到的 ICTCLAS(Java 64 位版)对其二次开发实现。处理过后的金融信息文本想要加入预测模型中，必须得到量化。而利用新闻中反映的信息强度可以作为良好的对接口，情感值的计算核心的内容就是为上一步骤获得的文本信息生成一个量化的值。此过程可以使用 Hownet 对新闻情感值进行处理。利用在之前获得的分词结果，在词典中进行匹配，确定各个词的量化值，有了每个词的量化值后，就可以得到整个文本的情感值。特别地，由于金融领域的特殊性，采用人工方法对词典进行了修改，基于路透金融词典对 Hownet 正负面词汇词典进行了修改，添加了 100 个正面词汇和 150 个负面词汇。

6.2.3 基于指标与维度的数据优化

在金融股市预测的研究中，如何选取正确的输入指标是问题的关键之一，也一直是研究的难点之一。到现在选择哪些指标作为模型的输入是最优方案都无定论，大部分的学者往往通过依靠历史经验或者主观的臆断进行决策，本节采用灰色关联度的算法进行指标筛选，挑选出对预测模型的建立具有显著性影响的指标。同时，采用流形算法对选取的指标进行降维处理，保留原始数据的特征，同时获取更低维的输入，进一步优化目标模型的预测能力。

同时，在股票预测的输入数据中，时间序列数据高噪声的特点严重影响着预测效果。本书引入了小波变换降噪的方法对股市时间序列进行处理，利用四种小波基函数和六种不同的阈值规则进行遍历寻优，具体数据的优化方法如下文所示。

1. 基于灰色关联度理论选取输入变量

与预测值不相关的变量可能导致机器学习算法效率的低下，对最终的预测结果造成负面的影响，因此通过对输入特征向量的选择，保留关联性强的输入量，删除关联性弱的输入量，能够有效提高机器学习算法的效率和性能，基于以上原因，可以考虑采用灰色关联度分析理论首先分析各输入变量与预测值之间的灰色关联度，选取相应的输入变量进入下一步的处理。

(1) 灰色关联度理论分析

在灰色关联度的计算方法中，利用斜率求取灰色关联度的方法分辨率较高，并且能够处理数据中的负数或零值。基于以上原因，众多研究采用了灰色斜率关联度方法，同样选取基于灰色斜率关联度，对输入指标进行灰色关联度分析，该方法具备对原始序列进行无量纲化数据变换时关联系数及关联度的值保持不变的优点，并且分析结果客观可靠。斜率灰色关联度分析方法的基本思想是利用因素序列曲线的平均相对变化情况的相似程度来计算灰色关联度。具体而言，就是将原始数据连接成折线，求出两个相邻元素之间的斜率，利用斜率判断出曲线的增减走势。最后将因素和参考因素之间的相对变化情况进行统计分析得出灰色关联度。

(2) 股票价格时间序列影响因素的改进灰色关联分析实现

如上所述，下面对股票价格时间序列求取灰色关联度。处理的对象基于表 6.5，将 $X_1 \sim X_{31}$ 作为因素序列，将次日收盘价作为参考因素序列，基于斜率的改进灰度关联法求取关联度大小。

该算法的具体思想和处理方法如下。

① 初值化处理

因素序列为

$$X_i = (x_i(1), x_i(2), \cdots, x_i(n)) \quad i = 1,2,\cdots,m$$

参考因素序列为

$$Y = (y(1), y(2), \cdots, y(n)) \quad i = 1, 2, \cdots, m$$

则因素序列和参考因素序列在区间$[k, k+1]$,$k=1,2,\cdots,n-1$ 上的斜率为:

$$\Delta x_i(k+1) = x_i(k+1) - x_i(k)$$

$$\Delta y(k+1) = y(k+1) - y(k)$$

求取因素序列和参考因素序列的均值为

$$\bar{x}_i = \frac{1}{m}\sum_{k=1}^{m} x_i(k)$$

$$\bar{y} = \frac{1}{m}\sum_{k=1}^{m} y(k)$$

求取因素序列和参考因素序列的 2 阶范数为

$$\text{Min}_{i,k} = \min\left(\left|\frac{1}{\bar{x}_i}\Delta x_i(k)\right|, \left|\frac{1}{\bar{y}}\Delta y(k)\right|\right)$$

$$\text{Max}_{i,k} = \max\left(\left|\frac{1}{\bar{x}_i}\Delta x_i(k)\right|, \left|\frac{1}{\bar{y}}\Delta y(k)\right|\right)$$

② 方向性处理

由于斜率有正负的区别,在实际的预测中,同向变动的因素序列能够给预测带来更大的帮助,参考肖新平设立的方向判别函数,设立判别函数 SGN,SGN 代表的含义如下:

$$\text{SGN}(\Delta x_i(k), \Delta y(k)) = \begin{cases} 1 & \Delta x_i(k) * \Delta y(k) \geqslant 0 \\ -1 & \Delta x_i(k) * \Delta y(k) < 0 \end{cases}$$

利用 SGN 函数能够将同向变化的数据转为正值,反向变化的数据转为正值,求和时同向变化居多的因素序列能够得到凸显。同时,由于考虑因素序列的同向变化比曲线的接近程度更重要,因此,在计算时引入:

$$1 - \frac{\text{Min}_{i,k+1} + \text{Min}_{i,k+1} + 1}{2\text{Max}_{i,k+1} + 1}$$

该参数满足 $0 \leqslant 1 - \frac{\text{Min}_{i,k+1} + \text{Min}_{i,k+1} + 1}{2\text{Max}_{i,k+1} + 1} \leqslant 1$,当斜率变化同向且数值相同时则该参数能够保障$\gamma_i(k)=1$。而斜率变化反向且数值相同时则该参数使得$\gamma_i(k)=-1$。当因素序列和参考序列斜率差距越大时,$1 - \frac{\text{Min}_{i,k+1} + \text{Max}_{i,k+1} + 1}{2\text{Max}_{i,k+1} + 1}$越趋近于 0.5,$\gamma_i(k)$会相应有效地降低斜率变化幅度对关联度运算的影响。当差距越小时则 $1 - \frac{\text{Min}_{i,k+1} + \text{Max}_{i,k+1} + 1}{2\text{Max}_{i,k+1} + 1}$越趋近于 0,与设计算法时斜率变化度越小的灰色关联度越小一致。综上所述,该参数可用于消减斜率绝对值差的影响,突出

因素序列的变化方向性。

因素序列和参考因素序列各斜率的灰度如式(6-15)所示：

$$\gamma_i(k)=\mathrm{SGN}(\Delta x_i(k),\Delta y(k)) * \frac{1}{1+\left(1-\frac{\mathrm{Min}_{i,k+1}+\mathrm{Max}_{i,k+1}+1}{2\mathrm{Max}_{i,k+1}+1}\right) * \left|\,|\Delta y(k)|-|\Delta x_i(k)|\,\right|} \tag{6-15}$$

因素序列和参考因素序列的总灰度如式(6-16)所示：

$$\gamma_i=\frac{1}{m-1}\sum_{k=1}^{m-1}\gamma_i(k) \tag{6-16}$$

从$\gamma_i(k)$的定义可以看出，$\gamma_i(k)$具有如下几点性质：

- $|\gamma_i(k)|\leqslant 1$。
- 其对称性、唯一性、可比性。
- $\gamma_i(k)$突出因素序列方向性的作用，变化同步越高，则灰度值越大，反之则越小。

方法的具体实现可以利用 MATLAB 软件，计算得到各输入变量和参考因素之间的灰色关联度，利用灰色关联度大小进行筛选，剔除值过小和值为负数的情况，得到的新的输入变量序列。

2. 流形算法数据降维

数据降维是数据挖掘研究中非常重要的一种工具和方法，其目的在于发掘高维数据中隐藏的内在结构，从而促使基于高维数据的分类、可视化和压缩得以更好地运行。如上文所述，在针对时间序列的数据挖掘中，降低输入向量的维数是一个重要也是必备的处理过程，因此，本书将引入流形理论作为对输入特征约简的处理手段。

(1) 流形学习理论

从数学意义上定义流形学习。给定数据集 $X=(x_1,x_2,\cdots,x_n)\subset R^D$，假设 X 中的元素可以通过低维空间中的集合 Y 利用某种非线性变化 f 得到，即$x_i=f(y_i)$，其中 $Y=(y_1,y_2,\cdots,y_n)\subset R^d$，$d\ll D$，$f:Y\rightarrow R^D$是一个光滑的嵌入映射。流形学习的目的就在于给出基于数据集合 X 的非线性映射$f^{-1}:R^D\rightarrow R^d$，获取这个高维空间到低维空间的映射结果 Y。

具体而言，数据降维就是通过线性或非线性的函数映射，将数据从高维空间映射到低维空间中，因而，数据降维方法可以分成两类：线性降维方法和非线性降维方法，其中，非线性降维方法就是通常所说的流形学习方法。具体情况如表 6.7 所示。

在表 6.7 中，主成分分析法(PCA)是使用最为广泛的线性降维方法，其基本思

想是在标准正交变换基础上方差较大的维视为主成分,其余为噪声。其优点在于具有最优线性重构误差,但是存在明显的缺点,主成分个数的确定没有明确的标准,同时不能用于处理非线性数据。同样地,线性判别分析和主成分分析十分类似,不过其主要针对分类问题,是一种监督方法,并不适用于处理较为复杂的问题。

表 6.7　数据降维方法一览表

<table>
<tr><td rowspan="2">线性方法</td><td colspan="2">主成分分析(PCA)</td></tr>
<tr><td colspan="2">线性判别分析(LDA)</td></tr>
<tr><td rowspan="6">非线性方法</td><td rowspan="4">保留局部性质</td><td>局部线性嵌入(LLE)</td></tr>
<tr><td>邻接图(Laplacian Eigenmaps)</td></tr>
<tr><td>Hessian 特征映射(HE)</td></tr>
<tr><td>局部切空间排列(LTSA)</td></tr>
<tr><td rowspan="2">保留全局性质</td><td>等距流形映射(Isomap)</td></tr>
<tr><td>多维尺度变换(MDS)</td></tr>
</table>

由于现实生活中的数据往往呈现高度的非线性结构,基于线性的方法存在局限性,无法揭示数据中复杂的真实规律,因此,基于非线性的降维方法逐渐在数据挖掘领域崭露头角。

在流形学习领域,2000 年 Tenenbaum 等和 Roweis 等学者在 Seienee 同时发表的两篇流形学习的文章,分别提出了等距流形映射(Isomap)和局部线性嵌入(LLE)算法。LLE 算法利用构造样本中各元素和它的邻域元素之间的一个重构权向量,将高维数据映射到一个全局低维坐标系中,保持邻域间的权值大小不变,这样的处理方法能够保留相邻点之间的集合结构,不仅能够有效地挖掘现有数据的非线性结构,同时保留了数据的平移、旋转等不变特性。Isomap 是基于多维尺度分析(MDS)的一种方法,它将样本中各元素和它的邻域元素之间的测地距离以两点之间的欧氏距离取代,而样本中各元素和邻域外的元素之间的测地距离则用流形上两点之间的最短路径来替换,这样能够更好地全局保留现有数据的几何结构。在此之后,流形学习成为新的研究热点,并获得了很大的发展。由于 Isomap 能够保持数据全局性质,很多学者利用 Isomap 算法对复杂的经济背景下的属性进行了约简等处理,在对 CSI300 股票聚类分类和对公司信用评价等具体问题取得了良好的效果,因此,本书引入 Isomap 算法对股票价格时间序列进行降维处理,以期降维后的新输入能够提高回归预测模型的精确度。

(2) 等距流形映射(Isomap)算法

① 构建邻域权重赋值图 G

设定输入样本为 $X=(x_1,x_2,\cdots,x_n)\subset R^D$,构建图 G 包含所有样本点。计算 X 中的每个样本点 x_i 和其余点之间的欧氏距离 $dx(i,j)$,如果 $dx(i,j)$小于域值 ε

或 j 是 i 最近的 K 个点之一时，可以认为它们是相邻的，此时设定图有边 $x_i x_j$，设定其权值为 $dx(i,j)$。

② 计算最短路径

如果点 i 和点 j 之间有边，则设定其初始最短距离 $dg(i,j)=dx(i,j)$，若不存在边，则设定 $dg(i,j)$ 为正无穷。设定 $l=1,2,\cdots,n$，$dg(i,j)=\min\{dg(i,j),dg(i,l)+dg(l,j)\}$，利用迪杰斯特拉算法构建最短路径矩阵，设定$D_g=\{d_g^2(i,j)\}$，矩阵包含了图 G 中任意两点 i、j 之间的最短距离的平方。

③ 计算 m 维嵌入

将 MDS 应用于矩阵 Dg，记 $S=\{s(i,j)\}=\left\{\delta(i,j)-\frac{1}{n}\right\}$，其中 $\delta(i,j)=\begin{cases}0 & i\neq j\\1 & i=j\end{cases}$。

$$H=\frac{-SD_gS}{2}$$

假设 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 是矩阵 H 的最大的 m 个特征值，$u_1,u_2,\cdots,u_m$ 是其对应的特征向量，$U=[u_1,u_2,\cdots,u_m]$，则 $T=\mathrm{diag}\left(\lambda_1^{\frac{1}{2}},\lambda_2^{\frac{1}{2}},\cdots,\lambda_m^{\frac{1}{2}}\right)$，$U^{\mathrm{T}}$ 即为 m 维嵌入结果。

（3）Isomap 算法实现

Isomap 算法可以基于 MATLAB 实现。从算法的具体实现中可以看出，和 Isomap 算法降维效果息息相关的有两个变量，一个是降维后保留的维数 m，另一个是选择邻域点 K 的大小。以下对两个参数的选取优化进行分析。

邻域点 K 如果过小，会产生将连续的流形划分为不相交的子流形的错误，对于全局性质的保留作用将会丧失，体现在结果上则表现为降维后数据点的丢失，但是如果 K 值太大，又会把整个数据集都变成局部邻域。一般而言，K 值的选取是根据主观确定的。可以利用增量的方法对 K 值进行检验，选取实验中能够保证对数据进行降维后不产生数据点丢失的 K 值。

对于 m 值的选取，根据先前学者的研究，嵌入的维数 m 应该满足限制条件 $1\leqslant m\leqslant m_{\max}=\log_2[N]$，其中，$N$ 是样本容量，利用增量式自寻优的方法对 m 值进行最优化选择，对于最优化的选取标准是利用 Isomap 降维后的输入进行预测时均方误差较小者为优。

3. 基于小波变换的阈值去噪方法

在对股票价格时间序列的研究中，由于股票价格时间序列具有不平稳、长尾性、含噪声等特点，使得许多传统的时间序列预测模型效果不好。故采用小波变化去噪的方法消除股票价格时间序列中的噪音，由此提高预测模型的性能。小波变换的降噪过程主要可以分为如下三步。

- 小波分解：选择小波基函数以及确定适当的分解层数 N 对原始信号进行

分解，得到低频系数(不含噪声)和含噪声的高频系数。

- 阈值降噪：设定适当的阈值函数和规则，对各层的小波系数进行降噪处理。
- 信号重构：将处理得到的高频系数与低频系数进行逆变换重构，即得到去噪的目标信号。

以沪深300指数时间序列为例，对小波消噪进行实证分析。从以下几个研究方面分别进行了实现。

(1) 小波分解

X 为原始数据信号，N 为分解层数，wname 为小波基函数，采用的小波基函数为四种 bior2.2、symN、dbN 和 haar，其中 N 代表消失矩，采用 db6 和 sym6，$[c, l]$ 是 X 的小波分解结构，c 由 $[ca_i, cd_1, cd_2, \cdots, cd_N]$ 组成，l 则存储了各矩阵的长度，分解函数如下：

$$[c, l] = \text{wavedec}(X, N, \text{wname})$$

获得各个尺度的细节系数的具体函数如下：

$$[cd_1, cd_2, cd_3] = \text{detcoef}(c, l, [1, 2, 3])$$

其中，cd_1、cd_2、cd_3 存储了各个分解层次的高频尺度系数。

$$ca_3 = \text{appcoef}(c, l, \text{wname}, 3)$$

ca_3 存储了低频系数。

(2) 阈值的获取

获得各序数矩阵之后，进行对阈值的计算。首先引入默认阈值的获取方法，其函数为 ddencmp，调用方式为：

$$[\text{THR}, \text{SORH}, \text{KEEPAPP}] = \text{ddencmp}('\text{den}', '\text{wv}', \text{X})$$

其中返回值 THR 表示计算得出的阈值，SORH 表示选择的阈值类型，分为软阈值(s)和硬阈值(h)两种，KEEPAPP 表示存储的低频信号 den 表示进行消噪处理，wv 表示选择小波基函数，X 表示原始信号。

运用 theselect，根据不同的阈值选择规则(TPTR)来计算确定信号 X 的阈值，具体示例如下：

$$\text{THR} = \text{thselect}(\text{X}, \text{TPTR})$$

特别地，引入小波方差分解软阈值规则(WVDSTR)

$$\begin{cases} \sigma^2 = \dfrac{\gamma_1{}^4}{\gamma_1{}^2 - \gamma_2{}^2} \\ \sigma^2 = \sqrt{2}\,\dfrac{\text{Median}(|C_{1,i}|)}{0.6745} \end{cases}$$

其中，σ 代表利用小波方差估计的噪声方差，r_1、r_2 分别代表第一、第二层小波系数的方差，Median($|C_{1,i}|$)代表第一层小波系数的中位数。当 $r_1 > r_2$ 时采用第一种方法进行估计，$r_1 < r_2$ 时采用第二种方法进行估计。

各层噪声标准差为式(6-16)：

$$\gamma_i = \gamma_1 * \left[1 - \frac{\gamma_1^2}{\sigma}\right]^{\frac{i-1}{2}}$$

如果阈值过大容易造成信号的失真，在综合考虑信号的预测能力后，取各层阈值的定义为式(6-7)：

$$\text{THR}_i = 2 * \gamma_1 * \left[1 - \frac{\gamma_1^2}{\sigma}\right]^{\frac{i-1}{2}} \tag{6-17}$$

(3) 阈值去噪实现和小波重构

该部分的实现利用函数 wdencmp 进行，具体操作方式如下：

[XC, CXC, LXC, PERF0, PERFL2]

= wdencmp('lvd', c, l, wname, N, THR, 'SORH')

其中，目标值 XC 就是所需经过消噪后重构的信号，lvd 则代表各层系数使用不同的阈值进行去噪，N 代表的含义如前，依然是小波分解的层数，THR 为确定的阈值序列，特别地，THR 矩阵的次数等于 N，PERF0 表示压缩率，PERFL2 代表信号质量。

(4) 去噪结构选择

如上文所述，在相同变量环境下采取的评价标准为信噪比和原始信号与去噪信号之间的标准差，信噪比越大越好，标准差越小越好。在不同变量环境下，将压缩率和信号质量也列入考查范围，若出现压缩率过小的情况则排除该方法。评价时先于同一层次同一阈值不同小波函数之间选优，再选取较优解进行比较，最终选取最优解。

(5) 小波变换去噪效果评价规则

一般而言，小波变换去噪效果评价规则取决于以下两点。

去噪后得到的信号应该和原始信号保有同等的光滑性，平滑度指标越小，去噪的效果就越好。

去噪后得到的信号和原始信号的信噪比越大，表明去噪效果越好；标准差越小，去噪效果越好。

6.2.4 金融预测模型及评价指标

1. 金融预测模型

当今，应用于股票预测的方法各种各样，主要涉及了局域的预测模型，全局的预测模型以及非线性方法等。其中，比较常用的方法有投资分析法、传统时间序列模型和机器学习方法等。这几年伴随着机器学习方法的快速发展，很多机器学习技术被应用到金融时间序列的预测中。本书重点介绍预测模型中较为典型的三种方法：自回归和移动平均(ARMA)模型、BP 神经网络模型(BPNN)、利用遗传算法优化参数的 SVM 算法(GA-SVM 模型)，分别对已处理的数据进行了预测。本

节将 ARMA 模型、BPNN 模型、GA-SVM 模型三者的实验结果进行分析对比。

此外，本文的预测模型是基于时间窗口 L，以 SVM 算法为例，在时间窗口 L 的时间范围内的数据作为训练集，$L+1$ 天的数据集作为测试集，利用基于遗传算法优化参数的支持向量机进行回归预测。预测模型中时间窗口 L 是一个变量，$3\leqslant L\leqslant 10$.。由于 L 范围的确定没有一个统一的标准，根据前人研究，L 的最大取值限定为 10，而考虑到 L 过小会导致训练模型不理想，因此限定最小值为 3。再对 L 进行增量计算，获取不同的 L 下的输出。输出则是次日收盘价。通过比较不同时间窗口 L 的评价指标的优劣，最终试验结果取最优 L 下的各项实验结果。

（1）利用遗传算法优化参数的 SVM 算法

SVM 是这几年来发展最快的机器学习方法之一，并且它已经被众多学者有效地运用于时间序列预测领域。尽管 SVM 已经在完成的研究中表现了较好的预测性能，但是它的预测性能和泛化能力经常会受到股票价格时间序列噪声和输入特征的影响，输入特征的高维度提高了预测模型的计算成本和过拟合的风险。在实践中，SVM 能够有效地解决小样本、非线性问题和高维模式识别问题。在股票价格预测的应用中，合理的输入特征和平稳的时间序列将会给预测准确性的提高带来良好的影响。

本部分的内容主要包括了利用 SVM 进行预测时的输入、输出、参数选择、预测性能评价指标的选择等。输入是基于之前利用灰色关联度分析、isomap 降维和小波变换去噪后确定预测模型的输入向量，时间窗口 L 是一个变量，$3\leqslant L\leqslant 10$，对 L 进行增量的计算，获取不同 L 下的输出。输出则是预测的次日收盘价。特别地，由于不同的核函数对于 SVM 的预测性能影响较大，根据之前的研究，径向基函数在股票价格时间序列的问题求解中能够得到较好的预测性能，因此可以选取径向基函数作为 SVM 核函数。在运用 SVM 进行回归预测时需要确定惩罚参数 C 和不敏感损失函数 ε。利用遗传算法（Genetic Algorithm）求解 SVM 最优参数，利用二者结合的 GA-SVM 模型求解回归问题。

GA-SVM 的预测评价指标最优表现如表 6.8 所示。

表 6.8　利用 GA-SVM 预测评价指标最优性能指标表

指标	RMSE	MAE	MAPE	TIC
指数	30.73552	22.96137	0.009472	0.006319
银行业	23.3692	16.741	0.00822	0.00579
民生银行	9.24068	6.47154	0.01052	0.00760
兴业银行	20.8688	14.0246	0.01038	0.00787
招商银行	16.1554	11.2530	0.00986	0.00715
中信银行	5.2621	3.9153	0.00976	0.00659

(2) 自回归和移动平均(ARMA)模型

ARMA 是一种自回归和移动平均模型,利用变量的历史数据进行对未来的预测。ARMA 模型一般需要考虑两个参数 p 和 q,其中 p 表示自我回归阶数,q 表示移动平均阶数,模型可表示为 ARMA(p, q)。在实际应用中,参数的确定一般按照 AIC 准则,即赤池信息准则,该项指标数值越小,说明模型的拟合程度越好。

利用 MATLAB 实现 ARMA 模型,其中设定移动预测步数为 1,时间窗口由 4 至 10 变动,p 和 q 的变动范围设置为 1～2,若 p、q 过大会产生无法生成模型的后果,得到的最优求解结果如表 6.9 所示。

(3) BP 神经网络

人工智能算法 Back Propagation(BP)神经网络是一种机器学习方法,根据 Kolmogorov 定理,已知三层 BP 神经网络能够在任意误差内逼近任意连续函数,因此书中采用三层 BP 网络模型。三层 BP 神经网络包含了输入层、隐层和输出层。其中,输入层的神经元个数为 41 个,和输入维度持平。隐层结点数没有固定的取值标准,因此设定隐结点数量为 1～15 之间,利用增量的方法对模型进行测试,以预测结果和实际值的均方误差为评价标准,选择最优的隐结点数量。以沪深 300 指数为例,其中 RMSE 指标随隐结点变动的情况如图 6.7 所示。

表 6.9　利用 ARMA 预测最优性能指标表

指标	RMSE	MAE	MAPE	TIC
指数	46.04093989	36.08553022	0.014882903	0.009462474
银行业	38.42965	28.39847	0.01405	0.0095
民生银行	15.14437	10.38291	0.01653	0.012401
兴业银行	34.81213	22.6069	0.016678	0.013191
招商银行	24.91968	18.4991	0.016572	0.0112
中信银行	7.36803	5.69348	0.01421	0.00922

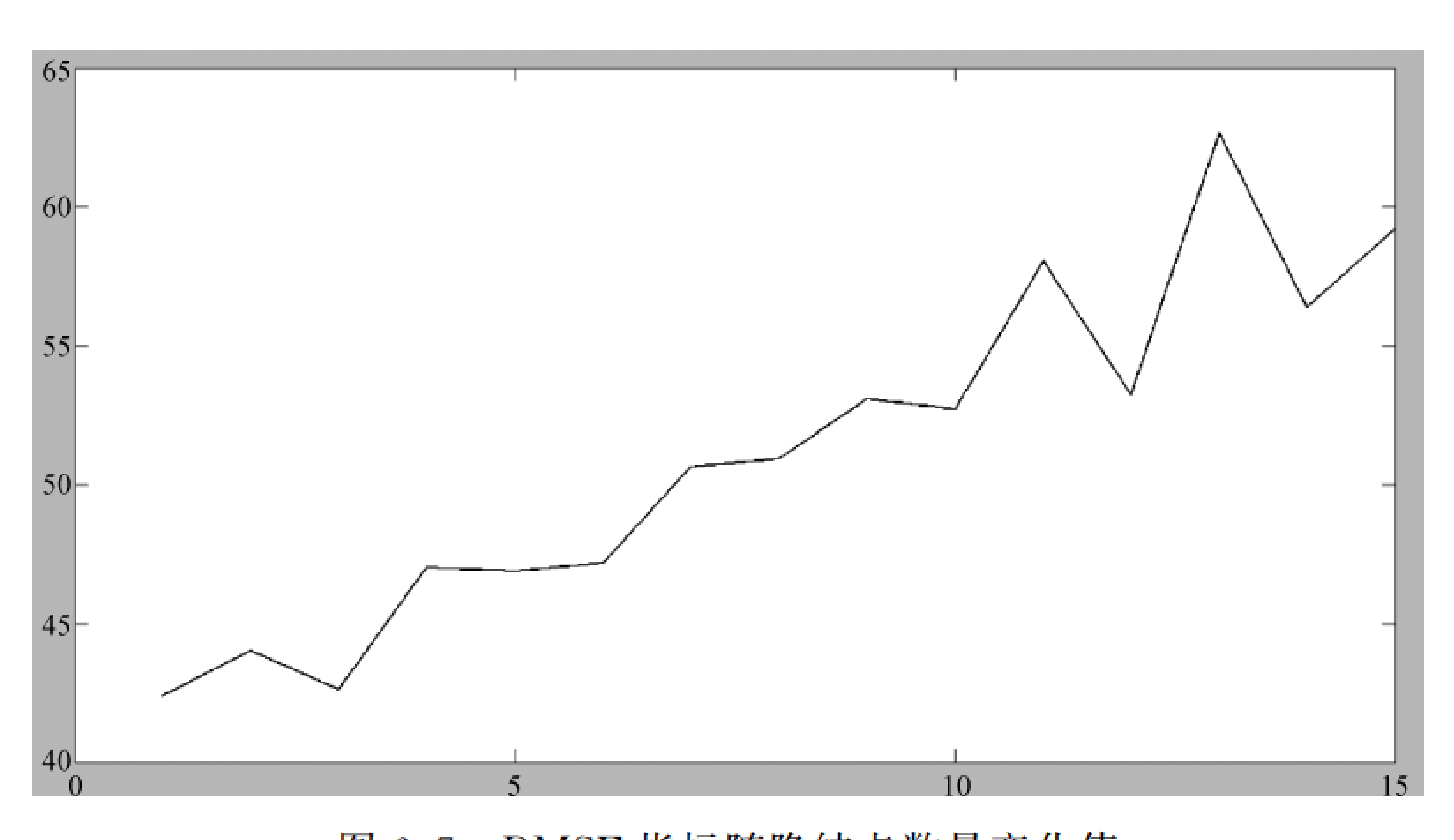


图 6.7　RMSE 指标随隐结点数量变化值

因此，选取的隐结点数为 1。BP 神经网络的预测评价指标最优表现如表 6.10 所示。

表 6.10　利用 BP 神经网络预测评价最优性能指标表

指标	RMSE	MAE	MAPE	TIC
指数	42.3965234	33.03305742	0.013633222	0.008724103
银行业	33.7569	24.3393	0.01199	0.00836
民生银行	13.2287	9.3899	0.01512	0.01089
兴业银行	29.9424	21.1707	0.01566	0.01129
招商银行	23.2052	16.2868	0.0144	0.01028
中信银行	7.16068	5.53374	0.01385	0.00896

综上所述，对时间序列预测研究中常用的 ARMA 模型、BP 神经网络及 GA-SVM 算法模型预测的最优性能指标对比得出，基于遗传算法求解 SVM 最优参数的 GA-SVM 模型在预测中表现最为理想。

2. 预测评价指标分析

选取以下四种性能指标用于对预测结果的检验，假设 PY_i 是根据预测模型得到的预测结果，Y_i 是对应的真实数据，四种评价指标定义如下：

（1）平均绝对误差（Mean Absolute Error，MAE）

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\mid PY_i-Y_i\mid$$

（2）平均绝对误差百分比（Mean Absolute Percent Error，MAPE）

$$\mathrm{MAPE}=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{PY_i-Y_i}{Y_i}\right|$$

（3）均方根误差（Root Mean Square Error，RMSE）

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(PY_i-Y_i)^2}$$

（4）希尔不等系数（Theil Inequality Coefficient，TIC）

$$\mathrm{TIC}=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(PY_i-Y_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(PY_i)^2}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i)^2}}$$

总而言之，以上四个指标中，RMSE 和 MAE 用于衡量预测值和原始值之间的误差大小，评价的标准应是越小越好。而 MAPE 和 TIC 一般用于衡量误差差别的程度，其二者的值均位于 0 和 1 之间，评价的标准是越接近 0 越好。如果出现结果相冲突的情况，则采用 RMSE 最低者为最优。

6.3 个性化服务

在过去的几年中，随着互联网的普及和计算机性能的飞速发展，计算机技术，尤其是社会计算和云计算也随之有了长足的进步。个性化推荐(Recommendation)技术，之前被广泛应用于电子商务网站，也渐渐地开始在社交网络上面崭露头角。

但是纵观各种类型的社交网站，无论是 Facebook 式的人人网，还是 Twitter 式的新浪微博，无论是在推荐的项目或者推荐的形式上面，都远远没有达到传统电子商务网站的高度。而反观需求方面，用户登录到社交网站，首先，好友推荐，而在好友推荐这一块，明显功能不够强大，系统甚至只能通过基于用户基本信息的固定算法来推荐；其次，新鲜事推荐，在我的众多好友中，一定有些人的新鲜事是我最愿意看到的，在众多言论中，一定有一些特定方面的信息是我所关注的，而在这一块，这些网站还基本处于真空阶段。

下面将介绍社交网站以及传统电子商务网站的个性化服务现状，以及用到的相关技术，最后会详细讨论一种实时的推荐算法。

6.3.1 国内社交网站推荐系统的发展现状

1. 人人网

人人网是由千橡集团对旗下著名的校内网更名而来的。人人网为整个中国互联网用户提供服务的 SNS 社交网站，给不同身份的人提供了一个互动交流平台，提高用户之间的交流效率，通过提供发布日志、保存相册、音乐视频等站内外资源分享等功能搭建了一个功能丰富高效的用户交流互动平台。

而人人网的推荐主要有三个方面，第一是广告和应用推荐，这一块跟人人网商业运作模式有关，基本为广告，不在本书讨论范围之内；第二就是好友推荐，用过之后很容易发现，这个推荐仅仅简单地基于共同好友和个人基本信息，按降序排列之后直接呈现给用户；第三是新鲜事推荐，其实还仅仅是一个雏形，甚至连推荐都没有，需要用户自己手动设置特别关注好友来实现(如图 6.8 所示)。

2. 新浪微博

新浪微博是一个由新浪网推出，提供微型博客服务的类 Twitter 网站。用户可以通过网页、WAP 页面、手机短信/彩信发布消息或上传图片。新浪可以把微博理解为“微型博客”或者“一句话博客”。您可以将您看到的、听到的、想到的事情写成一句话，或发一张图片，通过计算机或者手机随时随地分享给朋友。您的朋友可以第一时间看到你发表的信息，随时和您一起分享、讨论。您还可以关注您的朋友，即时看到朋友们发布的信息。



图 6.8　人人网好友推荐示例图

新浪微博提供的推荐服务则主要是围绕好友展开的，显然，相对于人人网的基本级别的推荐，新浪微博又要进了一步，第一，新浪微博可以直接从注册邮箱的联系人中帮你找寻好友（如图 6.9 所示），第二，也是最关键的一点，可以通过设置标签来让系统推荐拥有相同标签的好友，这一点，相对人人网来说，无疑是一个巨大的进步。

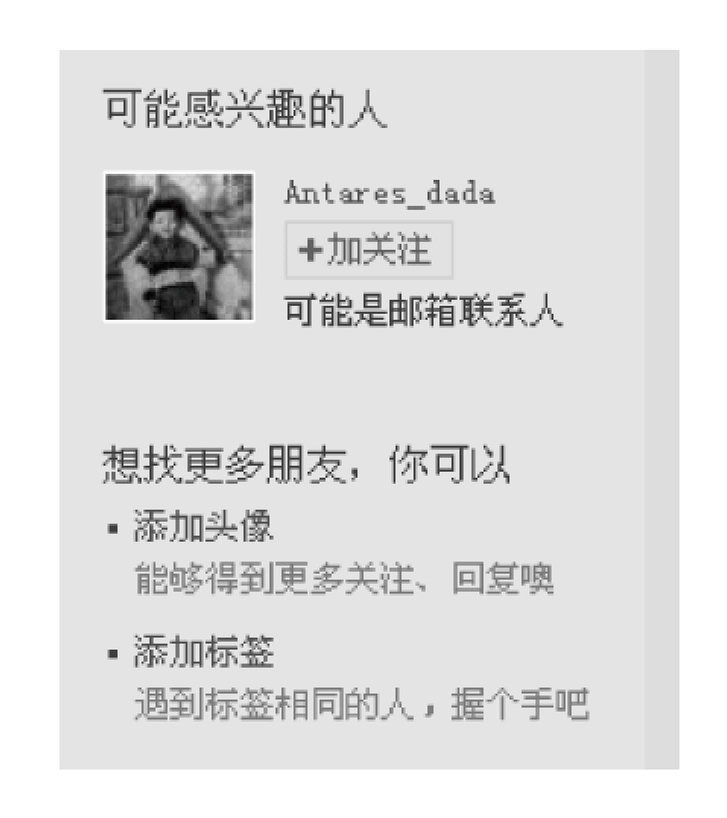


图 6.9　新浪微博好友推荐示例图

再来看看传统的电子商务网站的推荐系统，当用户买了一本书之后，会向你推荐其他的书籍，有基于协同过滤的（如图 6.10 所示），有基于用户行为的（如图 6.11 所示），还有基于商品基本属性的，不论是规模或是功能，都远远超过了社交网站的推荐系统，真正做到了“推荐用户真正关注的东西”。

总地来说，现在国内这些社交网站的推荐系统处于“聊胜于无”的状态，有简单的功能或者是模块，但是同传统电子商务网站的复杂推荐系统相比，则没有处于同一个高度。这就给了我们一个研究的课题：如何将传统电子商务网站的这些推荐技术融合到新兴的社交网站中，让以人为本的思想贯彻到底，关于这一点，将在后面的篇幅中阐明我们的观点和方法。



图 6.10 当当网推荐实例(Ⅰ)

购买本商品的顾客还买过



大学英语六级词汇21天速听速
¥13.10

- 大学英语四级词汇21天速
- GRE&GMAT阅读难句教程—
- GRE写作论证论据素材大
- GRE词汇逆序记忆小词典
- GRE作文大讲堂:方法.素
- GRE官方题库范文精讲—
- GRE词汇精选(便携版)—
- GRE数学高分快速突破—
- GRE写作高分速成-ISSUE

更多>

浏览本商品的顾客还看过

- GRE核心词汇考法精析—
- GRE阅读必备专业词汇—
- GRE 词根魔法图 (红牌黄



GRE词汇突破(升级版)/寄托蓝
¥35.20

- GRE词汇
- 词以类记:GRE词汇(附1c
- GRE巴朗词表—新东方大
- GRE考试官方指南(附CD-
- 17天搞定GRE单词——新
- GRE词汇精选:乱序版(

更多>

看过本商品的顾客会买

54% 会买本商品,还有:



23% 会买: GRE核心词汇考法精析—— 新东方大愚



8% 会买: 大学英语四级词汇21天速听速记(附光



6% 会买: 跟着名人名言背英语单词—3000词汇篇

浏览更多同类商品

GRE考试用书

- GRE考试官方指南词汇必备(
- 单词王·GRE单词速记
- GRE考试官方指南(附CD-RO
- GRE考前冲刺最新题库(配C
- GRE巴朗词表—新东方大愚
- 太傻单词——GRE BIBLE 叉
- GRE核心词汇考法精析(便

更多>

GRE考试

- GRE考试官方指南词汇必备(
- 单词王·GRE单词速记
- GRE考试官方指南(附CD-RO
- GRE考前冲刺最新题库(配C
- GRE巴朗词表—新东方大愚
- 太傻单词——GRE BIBLE 叉
- GRE核心词汇考法精析(便

更多>

图 6.11 当当网推荐实例(Ⅱ)

6.3.2　推荐的相关技术

1. 协同过滤

常用在电子商务的推荐系统里。

协同过滤推荐(Collaborative Filtering Recommendation)在信息过滤和信息系统中正迅速成为一项很受欢迎的技术。与传统的基于内容过滤直接分析内容进行推荐不同,协同过滤分析用户兴趣,在用户群中找到指定用户的相似(兴趣)用户,综合这些相似用户对某一信息的评价,形成系统对该指定用户对此信息的喜好程度预测。

协同过滤相对于传统的文本过滤,有一定优势,也有一些劣势。优势在于它能过滤一些难以对内容进行分析或者难以表达的东西(例如图片,音乐,质量等),同时能使推荐变得有新颖性。缺点也是明显的,由于用户对客体的评价非常稀疏,信息有限,使得结果可能不够精确,随着用户和客体的增多,系统效率会降低,同时,很有可能漏掉一些东西,使其永远得不到推荐。

协同过滤可以分为两种,基于用户的或者是基于项目的。

它一般采用最近邻技术,利用用户的历史喜好信息计算用户之间的距离,然后利用目标用户的“最近邻居”对商品评价的加权评价值来预测目标用户对特定商品的喜好程度,系统从而根据这一喜好程度来对目标用户进行推荐。

首先假设找到和此用户有相似兴趣的其他用户,则会对找到这个用户真正感兴趣的内容有一定帮助。所以,协同过滤的一般步骤为:交易数据库→测量用户间相似性→寻找相似用户→计算商品的购买可能性→根据购买可能性推荐商品。

2. 内容过滤

内容过滤是对网络内容进行监控,防止某些特定内容在网络上进行传输的技术。主要实现有软件和硬件两种。

当然这里指的是使用软件方式的内容过滤,基本原理就是根据客户的喜好和习惯跟内容进行对比。同样有很多优点,也有很多缺点,优点就是简单,有效,而缺点正是协同过滤的优点,对于一些难以对内容进行分析或者难以表达的东西,使用内容过滤是没有办法处理的。

3. 数据挖掘

数据挖掘是一种透过数理模式来分析企业内储存的大量资料,以找出不同的客户或市场划分,分析出消费者喜好和行为的方法。

数据挖掘可以做七种事情,分别是:分类(Classification),估值(Estimation),预言(Prediction),相关性分组或关联规则(Affinity Grouping or Association Rules),聚集(Clustering),描述和可视化(Description and Visualization),复杂数

据类型挖掘。

数据挖掘又分为直接数据挖掘和间接数据挖掘。

先来说一种比较常见的数据挖掘算法 Apriori 算法。

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里,所有支持度大于最小支持度的项集称为频繁项集,简称频集。

该算法的基本思想是:首先找出所有的频集,这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规则,这些规则必须满足最小支持度和最小可信度。然后使用第一步找到的频集产生期望的规则,产生只包含集合的项的所有规则,其中每一条规则的右部只有一项,这里采用的是中规则的定义。一旦这些规则被生成,那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集,使用了递推的方法。

6.3.3 一个例子:动态信息推荐

此模块为动态信息推荐,放到人人网上面来说可以简单地理解为新鲜事推荐,在这一个模块中,要解决一个非常重要的问题,就是消息的时效性的问题。我们知道,在好友推荐模块,任何好友的地位都是平等的,也就是说新注册的好友和原来注册的好友之间没有什么区别。但是对于动态信息来说,时效性却是非常关键的一项属性,举一个最为简单的例子,“9·11 事件”在 2001 年 9 月 12 日这一天绝对是世界上最为火爆得新闻,没有之一,但是,如果将其放到十年后的 2011 年,这显然不足以称之为新闻,这就是时效性,这一属性是之前好友推荐里面所没有的。

但是,在解决时效性之前,先来看看如果没有时效性这一属性的推荐方法,首先在这里需要说的一点,由于动态信息的推荐跟好友推荐有一个本质的区别:这个推荐程序会经常执行,在用户每次打开首页的时候,动态信息推荐都会执行一次,并将结果作为推荐内容呈现给用户,所以此处,应该尽量选择方法较为简单但是效率高的算法来减轻服务器的负担,而不是选择传统的,计算量非常巨大的“收割,分词,提词频”等耗时巨大的算法。所以通过综合考虑选择 Apriori 算法。

Apriori 算法是一种最有影响的挖掘关联规则的算法,该算法先挖掘出所有的频繁项集,然后由频繁项集产生关联规则,许多挖掘关联规则频繁项集的算法都是由它演变而来的,虽然也需要挖掘出关联规则以便进行页面推荐,但是 Aprior 算法并不适合进行基于数据库的关联规则挖掘。这是因为数据库中所包含的是序列数据,我们需要的规则也是有时间戳的,因为访问网页的时间是有先后顺序的,例如:P1→P2 和 P2→P1 具有不同含义,而 Apriori 算法则没有考虑到时间的先后对规则挖掘的影响,它只是反映出访问 P1 的用户也访问了 P2。

先引入一个序列模式和关联规则：

（1）访问新鲜事 P1 之后有 30％的用户访问了新鲜事 P3。

（2）访问新鲜事 P1 和 P3 之后有 35％的用户又访问了新鲜事 P5。

（3）在用户一次访问过程中同时访问新鲜事 P1、P3、P5 的概率为 16.1％。

（4）经过聚类分析之后发现 P1、P3、P5 属于同一个类。

根据所发现的类似模式和规则，可以进行页面推荐，例如，当一位用户访问了页面 P1 和 P3 之后，根据分析以往用户的访问模式所得到的信息便可以向当前用户推荐页面 P5。

直接使用“滑动窗口”的概念，这样可以直接剔除不影响结果的用户行为。例如，用户访问了 P1、P2 和 P5 三个页面，则 W ＝ < P1,P2,P5 >，其中 W 表示滑动窗口。在这里设置 W 的默认大小为 3，即只用记录用户最近访问的三个页面。这个值得选择基于以下两个考虑：(1)最新访问过的页面更能反映用户当前的兴趣所在；(2)由于考虑了页面被访问的先后顺序，而浏览顺序完全一致的情况并不是经常发生，尤其是当 W 的值大于 3 的时候。而且，在实际算法中，当 W＝3 无法找到完全匹配的结果时，可以动态将 W 递减，直到 W ＝ 0 或者找到匹配的结果。

下面来看看具体算法的过程，将以 W ＝ < P1,P2,P4 >为例说明如何得到推荐页面。

取 W ＝ < P1,P2,P4 >中的第一个页面 P1，然后在数据库中开始找寻所有在访问 P1 之后又访问了 P2 的用户数 N1，如果 N1 > 0，则记录这 N1 个用户，执行第二步。

在这 N1 个用户的集合中，再继续寻找在访问了 P2 之后访问了 P4 的用户数量 N2，得到的 N2 如果大于 0，则将结果集记录下来，执行下一步。

此时的 N2，就是所有按顺序访问了 P1，P2 和 P4 的用户，显然，可以得到一个集合 PN，PN 中包含了所有 N2 中的用户在访问 P1，P2 和 P4 之后所访问的下一个页面。由于原算法需要验证支持度和置信度，考虑到效率和待会还要进行时效性验证，所以直接以一个非常简单的算法取得结果：在 PN 中，直接取占有量最大的那个网站，例如在 PN 中，P3 被访问了 13 次，P5 被访问了 20 次，P6 被访问了 48 次，则直接选择 P6，如果出现占有率相同的现象，则直接选择时间上比较晚那一个页面。

根据以天津大学网站 2003-03-01～2003-03-07 一周的 Web 日志文件作为试验的对象，该日志文件共 105MB，经数据清理后剩余有效记录 378747 条。试验的目的是为了测试不同的支持度和置信度阈值对于推荐页面数量的影响以及滑动窗口 W 的大小设为 3 是否理想。试验发现，随着支持度和置信度阈值的提高，所得到推荐页面数量明显减少，同时，随着滑动窗口 W 的增大，得到的推荐页面数量也

呈减少的趋势。可见选择合适的支持度、置信度和滑动窗口 W 对于最后得到的推荐页面数量有明显的影响，如果取值太小虽然可以得到更多的页面，可是页面间的相关程度明显降低，取值太大可以得到相关度很高的页面，但是得到的页面数量太少。针对算法 Predictor，采用支持度 $S = 2\%$、置信度 $C = 40\%$ 以及 $W = 3$ 可以得到较好的结果。当然，这是文献中作者对算法执行效率的测试，可以直接简单的认为，窗口大小取 3 是比较合理的一种做法，在得到这个结论之后，便可以开展下面的工作。

如果我们不考虑时效性，那上面这个算法可以算一个比较好的算法，但是，消息如果不考虑时效性就会发生像上面列举的九一一式的笑话。在此，考虑到推荐结果并不需要太过于精确，可以在上述算法中做出一个非常小的改动即可。

选择“加权”的办法，加入时间的权重，显然，时间越靠近当前时间，权重应该占得更重，而越靠近当前时间，权重的增加幅度应该减少，可以只考虑 10 天之内的新鲜事，因为 10 天前的事情，可以因其时效性滞后而直接将其忽略，所以取 240 发生时间是可行的。因此，直接将时间放在最后权重公式的乘号后面。

当然，整体上也要做出一些修改。原因是由于加入了时效性的验证，必须将上面的窗口不停地往前面滑动，具体来说就是将得到的结果作为上述算法中最后一个浏览的页面，将窗口 W 向后移动一格，再进行一次计算。在每次推荐过程中，默认 $D = 5$，即执行 5 次上面的算法，每次将结果纳入 W，并同时将 W 向后滑动一格。显然，当 D 增大，权重应该减小，而随着 D 的减小，可以认为权重减小的幅度是一定的，所以，可以将这个值放在权值公式的乘号前面，所以权值定义如式(6-18)所示：

$$R_I = (D + 1 - I) * (240 - (T_N - T_R)) \tag{6-18}$$

上式中，I 为执行的次数，T_N 为当前的时间，T_R 为 R 创建时的时间，单位为小时。算法执行之后，选在 R 值最高的 3 个页面直接呈现给用户。来看一下数据库的实现(如表 6.11)所示。

表 6.11　图用户浏览记录数据表

	uid	newid	testtype	testid
□	4	1	1	1
□	4	2	1	2
□	4	1	2	3
□	3	1	1	4
□	3	2	1	5
□	3	1	2	6
□	3	2	2	7
*	(NULL)	(NULL)	(NULL)	(NULL)

由于 testid 是按照浏览的先后顺序插入数据库的，所以 testid 可以从侧面反映出用户浏览网页的先后顺序，testtype 是指是新鲜事还是状态。数据字典如表 6.12 所示。

表 6.12　表用户浏览记录数据字典

字段名称	是否主键	字段类型	中文描述
testid	是	INT	主键
uid	否	INT	用户号
newid	否	INT	消息序号
testtype	否	INT	消息类型

6.4　本章小结

社会计算是一个方兴未艾的发展迅猛的新领域，社会化媒体是社会计算的研究对象，社会化媒体的出现更是为许多传统领域开启了一扇融入更多交互元素的信息之门。在社会化媒体计算中，情感分析是一项非常重要的分析手段，本章首先以微博作为典型的社会化媒体平台，阐述了社会媒体下的情感分析概念、情感分析的文本处理过程，详细给出了情感倾向分析的分类模型，以及情感分类的评价指标。这套完整的流程描述清晰地展现了社会化媒体上情感分析的基本过程。其次，股票预测一直是金融预测领域的一项艰巨任务，本章从社会化媒体信息的全新视角出发，详细阐述了融入新闻舆情的金融预测的数据获取及量化处理方法；随后，从输入变量的选择、维度缩减、阈值去噪三个方面分析了金融股市预测的优化方法；进一步通过实验比较了 SVM 算法、ARMA 模型和 BP 神经网络三种金融预测模型，并给出四种金融预测的评价指标。最后，本章详细描述了在社交网站上如何实现个性化服务，介绍了推荐系统的核心技术，并给出了一个在社交网站上完成动态信息推荐的实例。

思考题

1. 试谈你对在社会化媒体下的文本挖掘的情感分析的理解。
2. 试搭建一个简单的微博情感分析系统。
3. 试谈你对基于流形学习的社会化媒体下的金融预测模型的理解。
4. 查找相关资料，整理分析股票预测的方法。
5. 思考在现有社交网站上如何更好地提供个性化服务，请列举一些个人的想法。

第 7 章

社会化媒体跨平台挖掘

本章学习目标

- 理解社会化媒体跨平台挖掘的意义
- 了解跨平台的用户识别

互联网上分布着各式各样的社会化媒体，人们在这些社会化媒体上讨论相同或相似的话题。因此，融合多个社会化媒体的社会行为数据，跨平台地进行各种分析，能为人们进行决策提供更完备的社会行为数据。跨平台数据挖掘即是使用某些技术、方式将各种各样的社会化媒体进行融合，从而为社会计算提供更丰富的社会数据和更完整的社会网络结构。跨平台社会化媒体的数据挖掘是指将来自多个不同社会化媒体的原始数据进行集中和融合，从而为社会计算提供全面的社会数据。

在互联网中，不同的社会化媒体有着不同的定位目标，它们分别对应用户不同的需求，包括沟通、分享等。例如，人人网用于博客撰写，分享心得等；新浪微博用于记录、分享生活点滴状态；微信用于即时沟通交流等。基于不同的需求，人们往往拥有多个社会化媒体账号。显然，在这些社会化媒体中，用户是它们之间进行集成的天然桥梁。通过将不同媒体间的属于同一个体的不同用户归类，就能自然而然地将这些社会化媒体进行融合。因此，跨平台社会化媒体的数据挖掘，其最根本的任务在于社会化媒体用户的识别。从一定程度上说，社会化媒体数据挖掘是社会化媒体的用户识别问题。如果能够跨平台实现平台间的用户匹配，将能为社会建模和分析提供更充分地社会行为数据，为知识发现和决策支持等应用提供更为全面的社会网络结构和更完整的用户信息内容（User Generated Content，UGC）和更充分的社会行为数据。

在基于社会化媒体的诸多研究领域中，跨平台数据挖掘的研究还处于刚刚起步阶段。现阶段所能查阅的文献资料与其他相关领域相比要少很多。虽然社会化媒体数据挖掘属于数据挖掘的领域，但是，它与传统的数据挖掘不太相同。跨平台社会化媒体数据挖掘的首要任务在于用户识别。在过去几十年里，为了识别不同的实体对象，各个领域都有实体识别的研究，如价格表、文献识别以及犯罪数据库等。跨平台用户的识别方法可以借鉴现有的实体识别方法。

由于用户是社会化媒体之间的天然连接纽带。因此，跨平台社会化媒体的数据挖掘，在很大程度上说，是社会化媒体的用户识别问题。也即，识别出多个社会化媒体中同属于同一个人的账号，进而通过这些账号将多个社会化媒体进行有机融合。社会化媒体为用户识别提供了许许多多的元数据，包括用户名、用户性别、年龄、所在地、头像、签名以及其发布的多媒体内容信息、内容信息发布时间、发布地点、发布来源等信息。鉴于不同的社会化媒体，人们所能获取的信息不同，且信息的疏密不同。因此，人们根据实际情况，使用一个或多个这些信息进行用户识别。

7.1　基于用户名的用户识别

在社会化媒体所贡献的所有信息中，用户名是唯一的所有用户都必须有的信息项。因此，人们在做用户匹配时，理论上可以依据用户名进行识别。人们在不同的社会化媒体上选择用户名时，往往会遵循一定的行为模式。从这些行为模式中进行数据挖掘，就能识别出不同社会化媒体上同属于一个人的用户。

根据用户名进行用户识别的问题可以定义为：假定已知某用户 I 在 n 个社会化媒体中的用户名为 $U=\{u_1, u_2, \cdots, u_n\}$，则给定另一个社会化媒体的用户名 c，判定 c 是否为属于用户 I。也即，基于用户名的用户识别可以用如下函数表示：

$$f(U,c)=\begin{cases}1 & c\text{ 属于用户 } I\\ 0 & \text{其他}\end{cases}$$

当人们使用某个用户名的时候，就可以通过抽取隐藏在用户名后的行为特征，并转化为数字特征，进而使用机器学习算法来判定。由于机器学习比较清晰；因此，在基于用户名的用户识别中，更多的研究重点主要集中在用户行为特征的抽取上。基于用户名的用户识别技术的整体框架如图 7.1 所示。

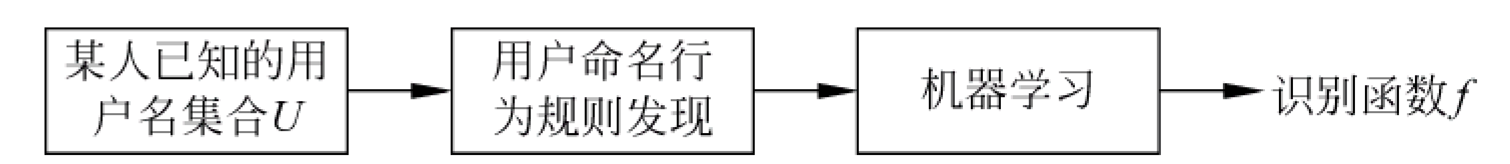


图 7.1　基于用户名的用户识别技术的整体框架

人们可以在不同的社会化媒体上选择完全不相关的用户名进行注册，从而使得研究人员无法从其用户名中获取足够的行为特征信息。然而，理论上只有满足最大熵原则的用户名才可能不会提供任何用户行为模式信息。也就是说，用户名长度是该社会化媒体所允许的最大长度，且用户名的所有字符都是完全随机的情况下，人们才无法从中获取用户行为规则信息。

很显然，受限于人自身的条件，满足最大熵的用户名几乎是不存在的(除非由于某些特殊的因素，由计算机自动生成用户名)。事实上，人短期内只能记住 7±2 个较长的字符；而且，人们对随机性字符的记忆能力较差，也即人们往往选择记住较为熟悉的字符串。因此，这些因素使得人们在选择用户名的时候，往往选用不是很长、非随机的字符串，该字符串中隐藏着人们在选择用户名时的潜在行为规则。从用户名中挖掘出用户名命名规则，就能在给定某自然人在某些社会化媒体中的用户名的情况下，判定另一社会化媒体中某个给定的用户名是否属于该自然人。通常情况下，用户命名规则的行为模式可以归纳为三类：人自身限制因素、外部因素和内部因素。本文以人自身受限因素为例，进行用户命名行为特征发现说明。

通常情况下，人们在选择用户名时，往往受记忆和知识的限制。这两方面的因素都可用于挖掘用户名命名特征。

7.1.1 记忆力受限因素

据数据统计，为了便于记忆用户名，59%的人习惯于在多个不同的社会化媒体使用相同的用户名。当用户名 c 在已知用户名集合 U 中出现时，c 和 U 有很大可能同属于一个人。因此，U 中 c 出现的次数可抽象为一个命名规则特征。当 c 在 U 中出现的次数越多，则 c 越有可能和 U 同属于一个人。当然，也有可能因为某个其他用户在该媒体中抢占了用户名 c，使得 c 和 U 不同属于一人。所以，c 出现在 U 中，并不一定意味着 c 和 U 就是同一个人。

另一方面，用户在构建用户名时，经常是从其常用的用户名中选取一个。这些可选的用户名的长度是不一样的。若用 l_c 表示用户名 c 的长度，l_u 表示已知用户名集合 U 中某用户名 u 的长度，则通常有 $\min(l_u) \leqslant l_c \leqslant \max(l_u)$。为了抽取用户名长度特征，通常用一个五元组来表示用户名集合 U 的用户名长度数字特征，即 $\{E[l_u], \sigma[l_u], \text{mid}[l_u], \min[l_u], \max[l_u]\}$，其中 $E[l_u]$, $\sigma[l_u]$, $\text{mid}[l_u]$, $\min[l_u]$, $\max[l_u]$ 分别表示 U 中用户名长度的均值、方差、中值、最小值和最大值。

7.1.2 知识受限因素

人们所掌握的任何一门语言的词汇量都是有限的。在有些情况下，人们所掌握的第一语言的词汇量要比第二、三语言要多一些。但其量一般都是一定的。正

如有人所统计的，人们所常用的英语词汇量在2000个左右。用户名通常是人们常用词汇的组合。因此，可以根据用户名所包含的词汇的个数来进行建模。其建模可借鉴用户名长度，建立一个五元组。

此外，任何一门语言，其字母或字的量是有限的，而人们所熟识的字或字母也是固定的。也就是说，人们往往习惯于选用他们所熟悉的字或字母来构建用户名。因此，也可以根据用户名中包含的字或字母的数量来构建类似的五元组。

通过上述模型特征的发现，获取用户进行用户名构建的基本特征，进而采用机器学习的方法，构建识别函数，最后，通过识别函数可以识别出给定的用户名是否与已知的用户名同属于一个人。

7.2　基于网络结构的用户识别

社会网络结构是社会化媒体中的一个重要信息。基于社会网络结构，人们可以进行各种各样的社会网络分析。通常情况下，社会网络分析是进行其他社会计算的基础。基于社会网络结构的用户识别可以定义为：给定两个社会化媒体的社会网络结构 $G1=\{V1, E1\}$ 和 $G2=\{V2, E2\}$，如何识别出尽可能多的用户匹配对(v_i, v_j)，使得 v_i 和 v_j 同属于一个人，其中 v_i 和 v_j 分别为社会化媒体 G1 和 G2 中的用户。

基于社会网络结构的用户识别方法通常分两步进行：种子结点识别和迭代识别。种子结点识别在于使用已有的稀疏属性或内容数据，挖掘出少量用户匹配对，进而使用这些少量已识别的用户迭代识别出更多的用户匹配对（如图7.2所示）。

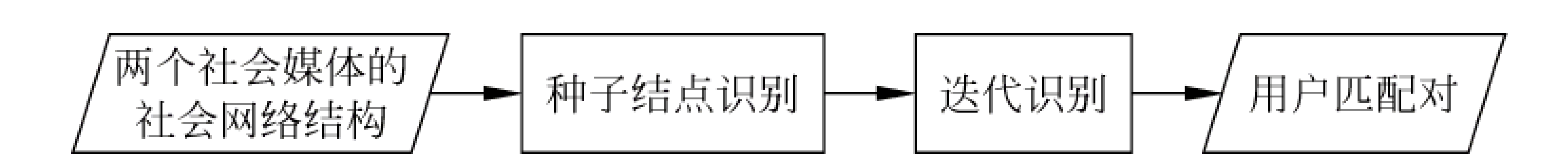


图7.2　基于网络结构的用户识别技术的基本流程

7.2.1　种子结点识别

在给定有限的属性下，只能根据所能获取的属性信息识别出少量的种子结点。由于不同的社会化媒体，其所能获取的信息不同；因此，现阶段并没有通用的种子结点获取方法。通常，人们根据所要进行数据挖掘的社会化媒体，有针对性地进行种子结点识别。由于种子结点识别不是基于社会网络结构的用户识别方法的核心，在某些情况下，种子结点将手工进行标注。

通常情况下，在不同的社会化媒体中，用户习惯于使用相同的用户名。因此，

基于用户名进行种子结点标注是一种方法。然而,在有些情况下,由于人们具有相似的行为习惯以及知识背景,因此,很多人会选用相同或相似的用户名,从而导致在两个社会化媒体中,相同用户名的用户并不一定同属于一个人。因此,往往需要加入额外的属性进行辅助判断。例如,在 QQ 微博和新浪微博的种子结点识别中,就可以使用用户名加签名的方法进行。即,当 QQ 微博中的某个用户和新浪微博中的某个用户具有相同的用户名和签名时,可认定这两个用户同属于一个人。

随着社会化媒体技术的发展,基于更好用户体验和商业意图等目的,许许多多的社会化媒体允许用户绑定其他社会化媒体的账户。例如,手机应用啪啪和唱吧等就可以绑定用户的新浪微博账户和 QQ 账户等。因此,通过账户绑定,可以准确、快速地获取种子结点。

此外,在某些社会化媒体中还可以通过用户属性中的网址(URL)信息进行种子结点识别。例如,通过分析 Twitter 账户中的 URL 属性,如果该属性中,包含其 Facebook 的主页信息,则能直接识别其对应的 Facebook 账户,从而实现 Twitter 和 Facebook 种子结点的识别。

7.2.2 迭代识别

给定两个社会化媒体的社会网络结构信息 $G1=\{V1, E1\}$,$G2=\{V2,E2\}$和已知部分给定的种子结点,进而识别出更多的用户匹配对是迭代识别的主要功能。迭代识别,顾名思义,就是根据已知种子结点,识别出部分的用户匹配对,进而将这些用户匹配对加入种子结点中,从而识别出更多用户匹配对。不断迭代上述过程,直到找出所有能识别的用户匹配对。最终,完成用户识别并进而完成跨平台社会化媒体数据挖掘。

在给定的两个社会网络结构中,迭代识别通常从某个社会网络结构中选取某个未识别的用户,进而通过某种算法计算出与该结点有共同已知邻接结点的另一个社会网络中的结点的匹配度。当匹配度大于已知设定的某个阈值时,则认为这两个结点为一个用户匹配对,同属于一个人。因此,在迭代识别过程中,最重要的因素在于如何计算两个未匹配结点的匹配度。

在两个基于有向图的社会网络结构的社会化媒体中,计算两个属于不同社会化媒体的用户之间的匹配度需要考虑如下因素:

(1) 边的有向性。由于社会网络结构是有向的,因此,在计算一对用户的匹配度时,可以通过计算两种匹配度,进而求和得到。其中,一种是基于入度的匹配度,另一种是基于出度的匹配度。将这两种匹配度相加,得到这两个用户最终的匹配度。

(2) 结点度。在社会网络中,用户的度往往服从幂律分布。因此,为避免因结

点度太大而导致识别准确率低，在使用结点度时，对结点度做平方根处理。

（3）离异度。为了提高用户识别的准确率，可以认为只有当来自第一个社会化媒体中的用户 u，同所有另一个社会化媒体中相关的用户的离异度大于某个设定的阈值时，才认为 u 很有可能同另一个社会化媒体中相关的用户中匹配度最高的用户相匹配。离异度的计算公式如下：

$$\frac{\max(X)-\max_2(X)}{\sigma(X)}$$

其中，max 和 $\max_2$ 分别为 u 同另一个社会化媒体中的用户匹配度的最大值和次大值，σ 为这组匹配度的标准差。只有当离异度大于某个设定阈值时，才认为 u 与最大匹配度的用户相匹配。

鉴于上述考虑，最终可以采用如下算法进行基于社会网络结构的用户识别。

步骤 1：遍历社会网络 G1 中所有未匹配的用户。

步骤 2：计算当前用户同所有与之有共同相邻用户的所有社会网络 G2 中的用户的匹配度值。

步骤 3：计算所获取匹配度值的离异值。

步骤 4：当离异值大于某个设定阈值时，则认为匹配度值最高的用户与当前用户同属于一个人。

匹配度的计算方法如下。

步骤 1：获取社会化媒体 G2 中与当前用户有共同指向已识别用户的用户。

步骤 2：当前用户同所获取的社会网络 G2 中各用户的出度匹配度等于 G2 中各用户的出度的平方根的倒数与这两个用户共同被指向的已识别的用户数的乘积。

步骤 3：获取社会化媒体 G2 中与当前用户有共同被指向已识别用户的用户。

步骤 4：当前用户同所获取的社会网络 G2 中各用户的入度匹配度等于 G2 中各用户的入度的平方根的倒数与该两用户共同指向的正识别用户数的乘积。

步骤 5：当前用户同步骤 1 和步骤 3 中所获取的社会网络 G2 中各用户的匹配度为其入度匹配度和出度匹配度的和。

7.3　本章小结

各式各样的社会化媒体平台丰富着人际之间的交流方式，每个平台都汇聚了人们生活的不同层面的大量数据，如何融合这些平台数据来挖掘用户的深层信息成为数据挖掘领域的新问题，而用户识别是跨平台社会媒体数据挖掘的首要任务。本章首先从人们在选择用户名的行为特征入手，给出了基于用户名的用户识别方法的探索；其次从网络结构的属性特征入手，阐述了种子结点迭代匹配的基于网

络结构的用户识别方法。

思考题

1. 根据你的理解，试谈谈跨平台的数据挖掘的难点和问题。
2. 思考在跨平台下如何实现用户识别的问题，谈谈你对这个问题的设想。

第8章

群 体 智 慧

本章学习目标

- 理解群体智慧在社会计算中的意义和含义
- 理解五种典型的群体智慧的算法思想

弗朗西斯·高尔顿(Francis Galton，1822—1911)是英国优生学家、心理学家，差异心理学之父，心理测量学上生理计量法的创始人。1906年秋天的某一天，他来到一个乡村集市参加一年一度的英格兰西部食用家畜和家禽展览会。这个展会是当地居民组织的对彼此饲养的牛、羊、鸡、马和猪等家禽家畜的品质进行评论的集市。当然，高尔顿来此的目的不是参加对这些动物的大众评论，而是一方面希望对于家畜的体质进行评估，另一方面希望从中发掘和倡导好的饲养方法。

高尔顿在会场漫步的时候，意外地被一处竞猜公牛重量赢大奖的地方所吸引。人们需要对一头肥壮的公牛进行鉴赏，同时需要给出自己对于公牛宰杀和去毛后的重量的估测。估测最为接近的人将赢取大奖。人们只需要花6便士就可以进行一次竞猜。这项竞猜赢来了各种各样的一共800个人来碰运气。他们来自各行各业，包括对此很在行的农民和屠夫，同时也包括对此一窍不通的其他行业的人。

高尔顿对整个竞猜过程产生了极大的兴趣。当竞猜结束之后他对所有参加竞猜打赌的人的估测进行了一系列的统计分析。他最终得到了787份有效的猜测结果。他计算了所有竞猜者竞猜数据的平均值。从某种程度上说，这个平均值实际上就是所有参加人员通过集体智慧对于公牛重量这一问题的集体抉择。

很显然，除农民和屠夫外的其他人毫无经验，只是凭借自己的想象和旁人的指引才做出估测的，这种估测理应是非常不准确的。换句话说，如果将整个竞猜过程看作是一个集体的决策过程，那么，农民和屠夫的观点应该是最具价值的，而其他

人的观点则会构成最终的决策噪音，使其产生偏差。基于以上的分析，高尔顿认为这个平均猜测值与实际值会相去甚远。不过令人吃惊的是，他错了！这个混合着内行人和外行人猜测的平均值是1197磅，而实际上这头牛的净重是1198磅。因而可以说这个集体的判断称得上是完美的。高尔顿后来由此写道："群体对于民主判断的准确性要比预想的可信得多"。

以上竞猜者对于公牛体重的猜测可以说是群体智慧行为的一个完美的诠释。群体智慧是一种由某一群体共享的智能，它从多个个体的合作与竞争以完成某一共同任务的过程中体现。集体智慧在微生物、动物、人类以及计算机网路中均可以形成。群体智慧又被称为共生智能。通过对于这种群体智慧的研究，研究者们还提出了诸多通过模拟这类行为来解决优化问题的有效算法，包括蚁群算法、粒子群算法、人工鱼群算法、人工免疫算法等。从更广泛的角度来说，这类利用群体智慧的算法实际上是社会计算的特殊形式，即通过社会个体的合作或竞争完成某一特定的任务。本章将对这些算法进行描述，同时给出一些群体智慧在社会计算中的应用实例。另外，本章末尾的补充材料还给出了美国海军利用群体智慧成功寻找潜艇"天蝎号"的例子。

8.1 蚁群算法

蚁群算法(Ant Colony Optimization，ACO)，又称蚂蚁算法，是由马克·多瑞格(Marco Dorigo)于1992年在他的博士论文中提出的。蚁群算法通过模拟蚁群觅食过程中发现路径的过程来对给定的问题进行优化。

生物学家发现蚂蚁的觅食行为是一个相互协作的过程：蚂蚁在爬行过程中将沿爬行路径不断释放一种称之为"信息素"的化学物质，这种化学物质能够被其他的蚂蚁所感知，同时其浓度也能够被识别。当一只蚂蚁在行进过程中感知到这种信息素时，它将倾向于沿着已有信息素的路径前进。当蚂蚁感知到有多条路径同时都存在信息素的时候，它倾向于向信息素浓度较高的路径前行。在这种系统行为的驱使下，由于经由最短路径达到的时间是最短的，导致随着时间的推移，经由最短路径到达觅食地点的蚂蚁数量比其他的路径要多，从而在这条路径上的信息素浓度也较高，更多的蚂蚁愿意从这条道路经过。同时，那些不常被经过的路径，由于信息素的挥发而浓度变低，愈发不会有蚂蚁经过。随着这种效果的不断地累计和放大，即便在最开始时蚁群通过各条路径前往食物源地点的概率分布是完全随机的，最终所有的蚂蚁也都会从距离最近的路径经过。图8.1是这种行为的一个简单示意图。

受此启发而得到的蚁群算法的优化过程由以下三个机制组成。



图 8.1 蚂蚁觅食过程

（资料来源：http://upload.wikimedia.org/wikipedia/commons/thumb/3/34/Safari_ants.jpg/440px-Safari_ants.jpg）

（1）选择机制：信息素浓度越大的路径被选择的概率相对越大。

（2）路径更新机制：某一路径上的信息素浓度会由于某一只蚂蚁的经过而得到一定程度的增强；同时，随着时间的推移，信息素会以一定的速率挥发而使得浓度降低。

（3）协调机制：蚂蚁间是通过感知路径上信息素的浓度来进行相互通信和协同工作的。

蚁群算法的核心思想在于，工作于蚁群中的某一个蚂蚁个体是不需要获知整个系统的信息的，事实上这也是不可能的。它仅仅需要获取它自身周围的信息，这可能是与周边的某些其他蚂蚁的交流抑或是感知它周边各条路径的信息素分布等。根据这些局部信息，蚂蚁个体将进行局部的调整或优化。随着这样的局部调整的进行，整个系统也随之逐渐进入更加优化的状态。实际上，这和我们熟知的社会计算非常类似。处于社会中的个体无法获取整个社会的信息，而只能依据自己已知的局部信息进行自我的调整，而这种调整实际上即是对整个工作的优化。蚁群算法中，规定蚂蚁个体自我局部行为调整的准则主要包括以下几点。

（1）工作环境。在一般算法实现中，蚁群的活动区域是一个被划分成具有多个小方格的方格世界。任意一只蚂蚁都处在这样的一个虚拟的环境之中，它所面对的对象包括障碍物、其他的蚂蚁、信息素、食物源、巢穴。其中，信息素分别产生于回家过程中和觅食过程中。

（2）观察范围。蚂蚁个体具有一个有限的观察范围。一般来说用一个参数 r 来设定其观察的区域 Θ。例如，$r=3$，则蚂蚁的观察范围是以它当前位置为中心的 3×3 的方格区域，并且其单步移动的范围也在这个之内。

(3) 觅食(巢)规则。某一只蚂蚁在当前所处的位置查看自己的观察范围 Θ 内是否存在食物,如果存在的话则直接移动过去;否则,查看自己的观察范围 Θ 内的哪一点具有最高浓度的信息素,同时向那一点移动。蚂蚁允许以一个极小的概率犯错,即在这一概率条件允许下移向的不是信息素浓度最高的地点。蚂蚁在寻找巢穴的时候使用相同的规则。

(4) 移动规则。在其观察范围之内,某只蚂蚁首先会选择移动到食物源(或巢穴);如果不存在,则移动到信息素浓度最高的区域;如果观察区域内无信息素,则依照原来的移动方向继续移动。

(5) 避障规则。如果蚂蚁在决定移动方向之后发现线路上有障碍物,则沿障碍物随机选一个方向绕开,如果此时选择绕开的路径包含信息素,则遵循觅食(巢)规则选择避障路径。

(6) 信息素散播规则。蚂蚁在行进过程中不断散播对应种类的信息素。在寻找食物源(巢穴)的过程中释放去往巢穴(食物源)的信息素。并且,随着与巢穴(食物源)距离的增加,散播的相应信息素量减少。

在以上规则的约束下,离开巢穴找到食物的蚂蚁可以标记回巢穴的路径,同时从食物源回巢穴的蚂蚁会标注去往食物源的路径。随着经过的蚂蚁的数量的增加,这些路径会随之加强。图 8.2 是一个蚁群算法中蚂蚁觅食过程中寻找最短路径的示意图。

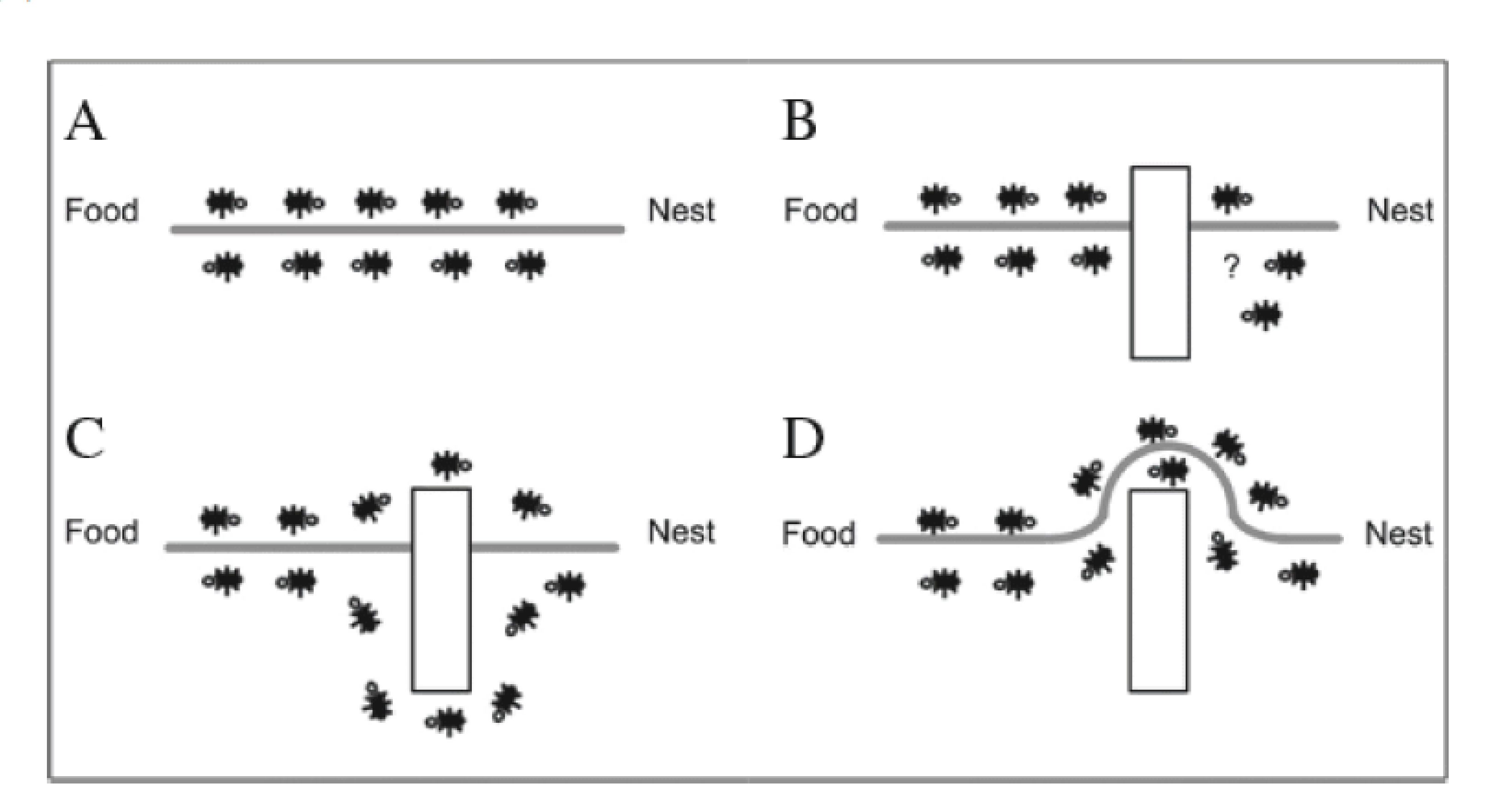


图 8.2 蚂蚁觅食过程寻找最短路径的示意图

(资料来源:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.4534&rep=rep1&type=pdf)

除了基本的蚁群算法之外,研究者们还开发了许多相关的扩展方法。其中包括精英蚁群系统、最大—最小蚁群系统(MMAS)、基于排序的蚁群系统(ASrank)、连续正交蚁群算法(COAC)、递归蚁群优化等。蚁群算法已被广泛地运用到各个

复杂的优化问题中。例如调度问题中的作业车间调度问题、集团车间调度问题、资源约束项目调度问题等；车辆路径问题中的容量限制的车辆路径问题、多分发点的车辆路径问题、随机车辆路径问题、时间窗的车辆路径问题等；集合问题中的集合覆盖问题、划分问题等；其他多类问题如数据挖掘、图像处理、旅行商问题等。

8.2　粒子群算法

粒子群算法(Particle Swarm Optimization，PSO)，又称粒子群优化，是由肯尼迪(Kennedy)和埃伯哈特(Eberhart)于1995年提出的一种著名的人工智能算法。该方法能够有效解决优化问题，从而一经提出就迅速得到了广泛的研究和应用。

粒子群算法通过模拟鸟类族群觅食时相互之间的信息传递机制来达到问题的优化目的。假想在一个空间里，所有的鸟都在搜寻食物。然而仅有一个地点有食物，而且没有一只鸟知道食物的具体位置是哪里，只知道自己与食物的距离。那么此时鸟群的策略可以是朝离与食物距离最近的那只鸟飞行。同时，每隔固定的一段时间，各只鸟重新汇报自己与食物的距离，以供鸟群中的鸟重新调整飞行的方向。在这个策略下，总体上来说，鸟群与食物的距离是越来越近的。因而，在一段时间过后，鸟群就能够找到食物的位置。粒子群算法的基本思想亦是如此，算法将每一只鸟当作粒子群系统中的单一粒子，赋予这些粒子以记忆能力，存储本身的飞行信息。同时，这些粒子能够与粒子群之中的其他粒子进行信息的交互。交互信息包括当前的飞行信息以及历史信息。根据彼此的信息，粒子从而调整自己的飞行方向和速率(即速度)。通过这种不断地交流和调整，算法期望粒子群中的某一个或者多个粒子能够寻找到问题的最优解。

粒子群算法的基本流程可以简述为：(随机或其他方法)初始化一个待优化问题的解集合(即多个解，亦即解空间中的多个点)，每个解被看作是粒子群优化算法的一个粒子。这些粒子以不同的速度(方向和速率)在解空间中运动。实际上，该速度是由该粒子本身的历史最优位置和整个粒子群所到达的最优位置共同决定的。即更新某一个粒子某一维度上的速度的式(8-1)是

$$v_{i,d} \leftarrow \omega v_{i,d} + \phi_p r_p (p_{i,d} - x_{i,d}) + \phi_g r_g (g_d - x_{i,d}) \tag{8-1}$$

其中，i 表示第 i 个粒子，d 表示粒子的第 d 维，v 表示对应的速度，p 表示该粒子历史最优位置，g 表示整个粒子群历史最优位置，x 表示该粒子该维度当前位置。r_p，r_g是随机生成的两个0～1之间均匀分布的随机数，用以调节粒子自身历史最优和粒子群历史最优对当前速度改变的影响。ω，ϕ_p，ϕ_g是由用户选择的用来调节算法效率和效果的参数，实际运用时需要不断调优。当更新完某一个粒子的所有维度上的速度之后，该粒子的自身位置将作对应的改变：

$$x_i \leftarrow x_i + v_i$$

图 8.3 给出了更新某一个粒子速度的示意图。其中，p^k 和 p^{k+1} 分别为粒子当前位置和更新后位置；v_{pbest} 和 v_{gbest} 分别为根据该粒子历史最优位置和粒子群历史最优位置求得的速度，即 $(p_i - x_i)$ 和 $(g - x_i)$；v^k 和 v^{k+1} 分别为更新前和更新后的速度；ω, ϕ_p, ϕ_g 是参数。

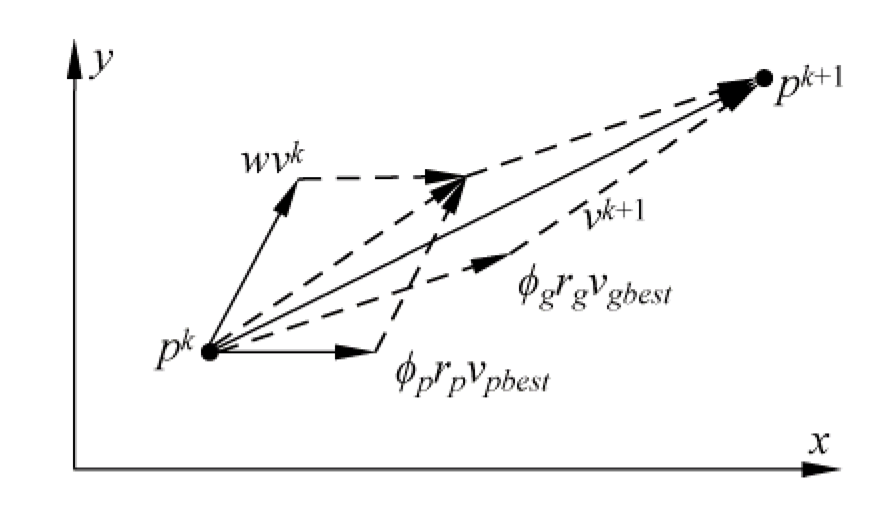


图 8.3　二维空间中，粒子群算法更新某粒子速度的示意图

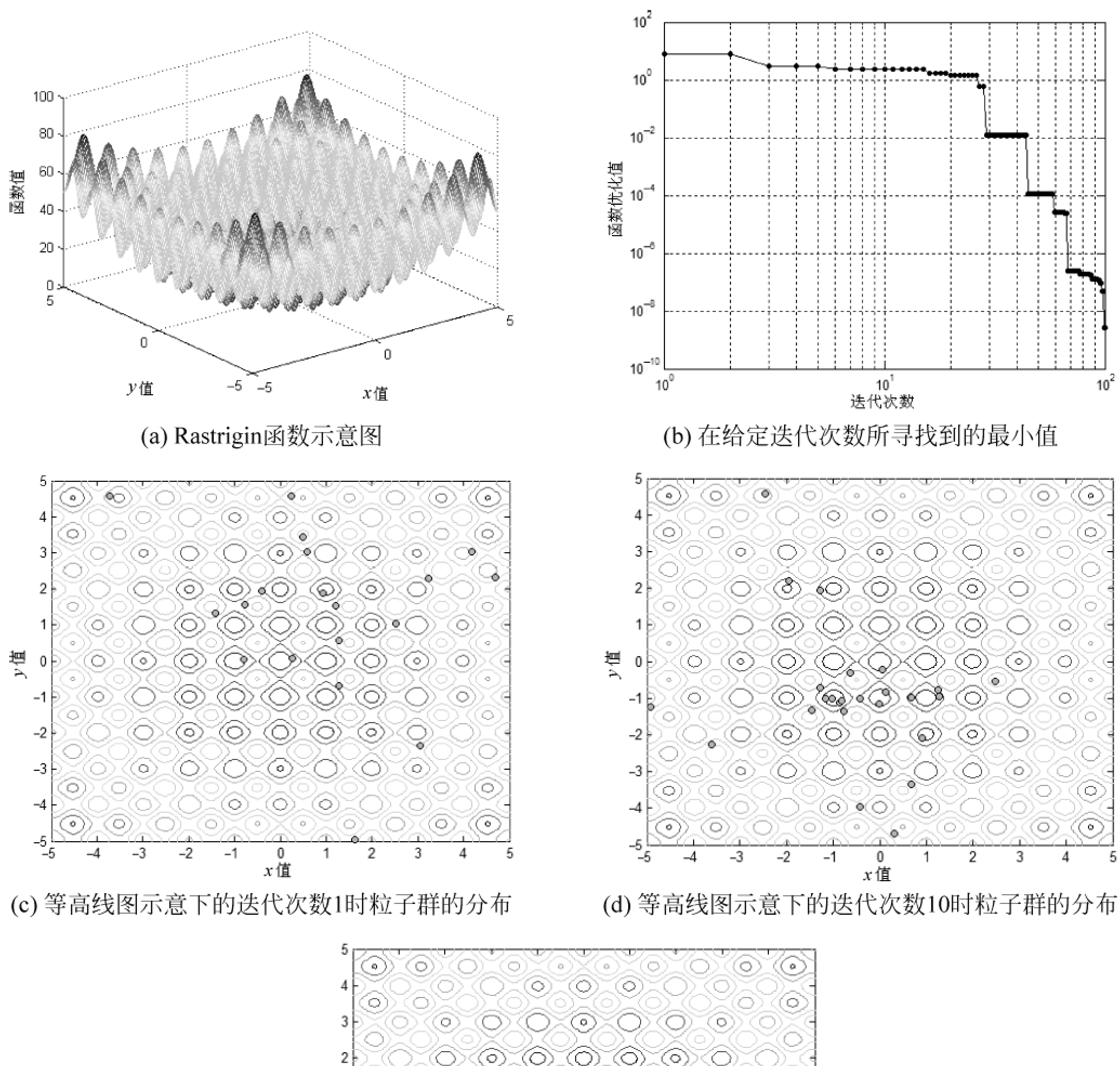
当更新完某一粒子的位置之后需要计算该位置上对应解的优劣，如果该解优于该粒子的自身历史最优解，则更新历史最优解。当更新完所有粒子的位置和历史最优解后，将所有的最优解与粒子群的历史最优解比较。如果存在优于粒子群历史最优解的粒子，则更新粒子群的历史最优解。重复以上的更新过程，直达算法收敛。算法运行结束后，g 即算法本次求解得到的最优解。

图 8.4 与图 8.5 所示的是利用粒子群算法进行函数优化的实例。其中，图 8.4 和图 8.5 分别是 Rastrigin 函数和 Ripple 函数，其函数定义分别如式(8-2)、式(8-3)所示。

$$f_1(x) = \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i) + 10), \quad -5.12 \leqslant x_i \leqslant 5.12 \tag{8-2}$$

$$f_2(x) = 0.5 + \frac{\left(\sin\sqrt{\sum_{i=1}^{n} x_i^2}\right)^2 - 0.5}{\left(1 + 0.001\sum_{i=1}^{n} x_i^2\right)^2}, \quad -10 \leqslant x_i \leqslant 10 \tag{8-3}$$

Rastrigin 和 Ripple 函数在 $\{x_i = 0, i = 1, 2, \cdots, n\}$ 处具有全局最优解，同时具有多个局部最优解。在图 8.4 和图 8.5 中，(a)给出了两维 Rastrigin 和 Ripple 的函数示意图；(b)给出了在迭代的过程中，所获得的函数值的变化过程；(c)～(e)分别给出了在算法进行到第 1 次迭代、10 次迭代和 100 次迭代时的粒子群分布情况。由图可以看出，粒子群算法的优化结果较好，均找到了函数的全局最小值。在迭代开始阶段，粒子群的分布较为扩散，几乎遍布了解空间的各个位置。随着迭代次数的增加，粒子群逐渐收敛到同一个位置。



(a) Rastrigin函数示意图

(b) 在给定迭代次数所寻找到的最小值

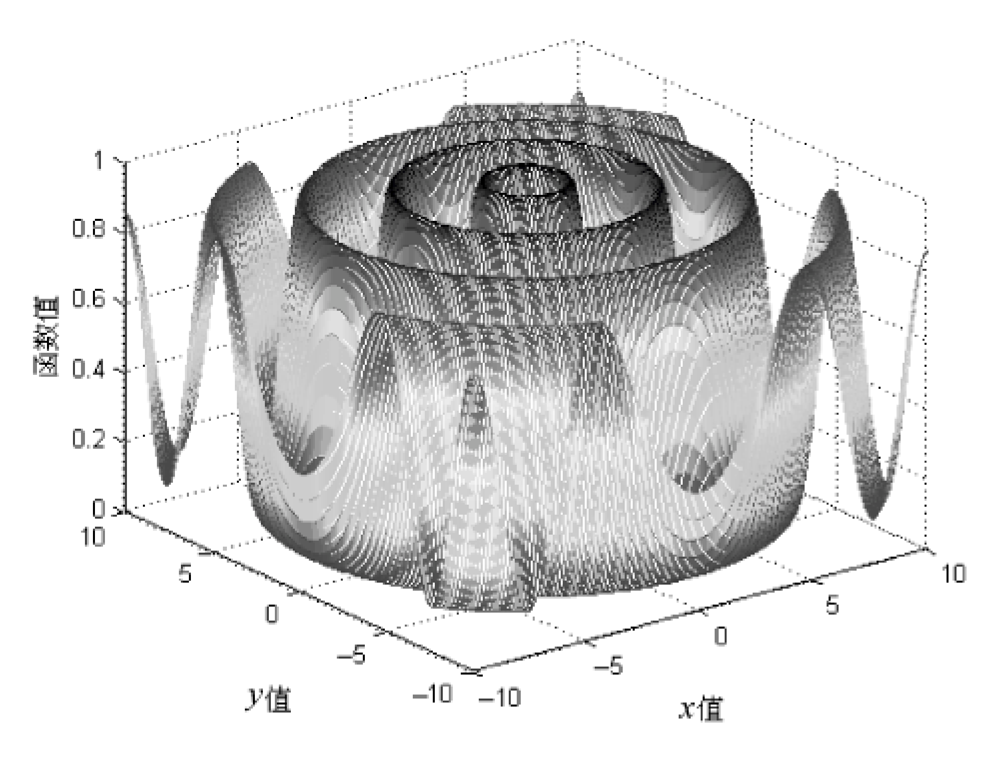
(c) 等高线图示意下的迭代次数1时粒子群的分布

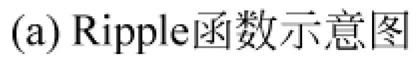
(d) 等高线图示意下的迭代次数10时粒子群的分布

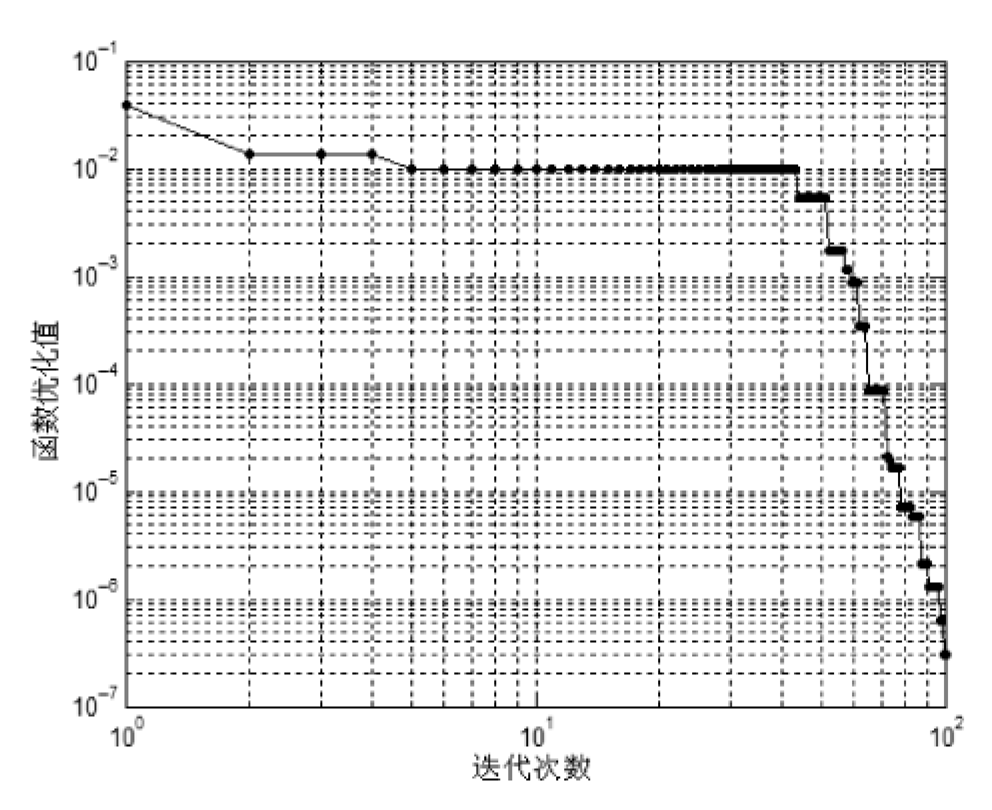
(e) 等高线图示意下的迭代次数100时粒子群的分布

图 8.4　粒子群算法寻找 Rastrigin 函数的最优值

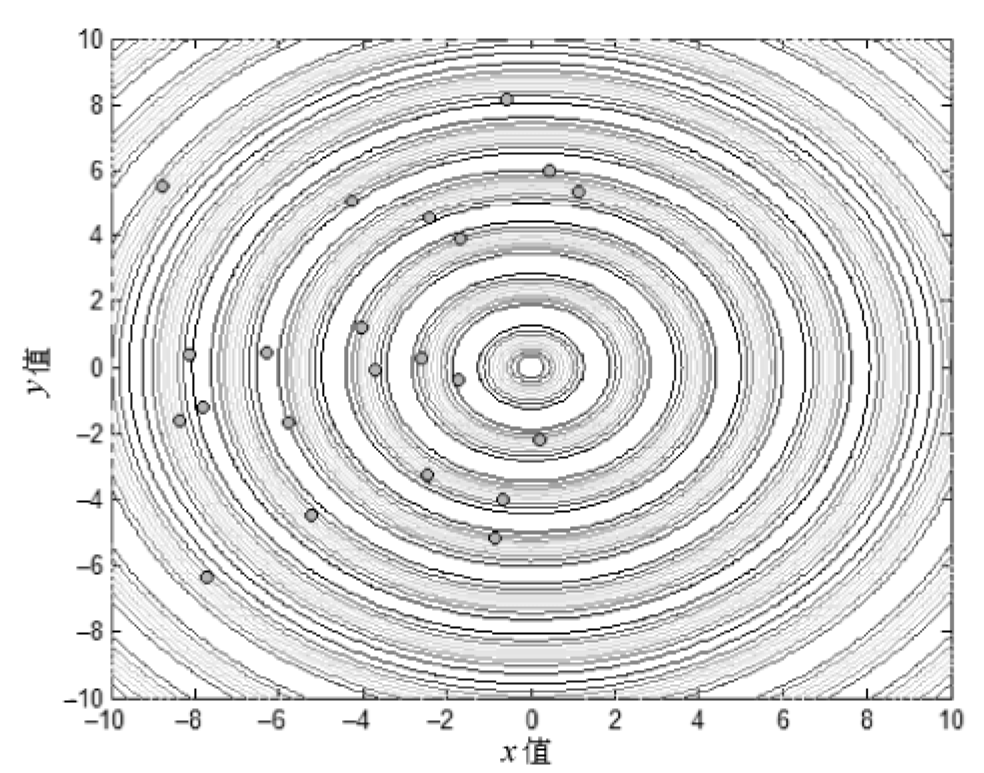
（c，d，e 图中黑色圆点为粒子）



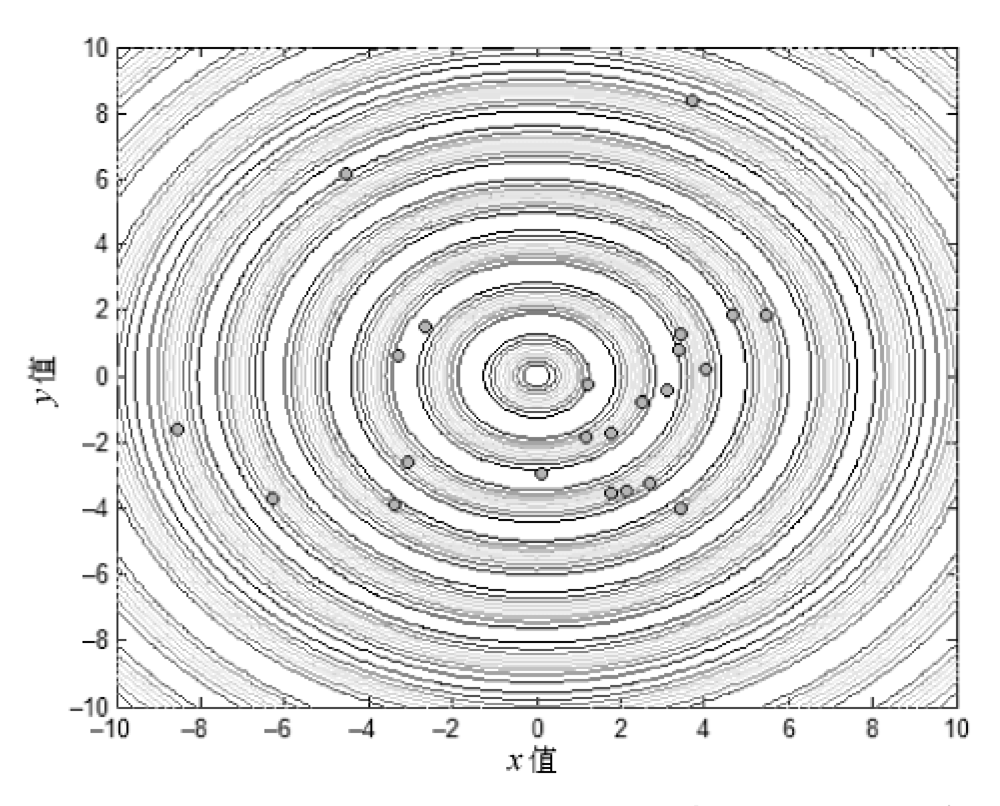

(a) Ripple函数示意图



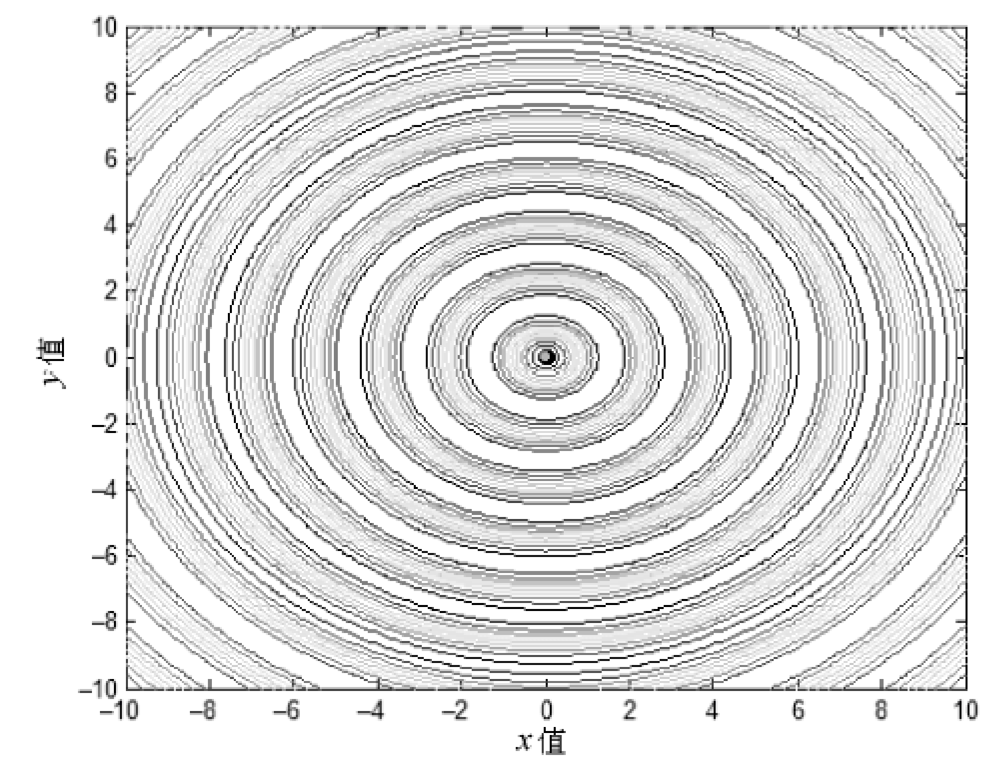
(b) 在给定迭代次数所寻找到的最小值



(c) 等高线图示意下的迭代次数1时粒子群的分布



(d) 等高线图示意下的迭代次数10时粒子群的分布



(e) 等高线图示意下的迭代次数100时粒子群的分布

图 8.5　粒子群算法寻找 Ripple 函数的最优值

（c，d，e 图中黑色圆点为粒子）

8.3　人工鱼群算法

人工鱼群算法(Artificial Fish Swarm Algorithm，AFSA)是李晓磊等(2002)提出的一种智能算法。与蚁群算法类似，人工鱼群算法通过模拟鱼群的群体觅食行为来进行优化。该算法假设水域中含有营养物质最多的地方是鱼群聚集的地方(鱼的生存密度最大)。

人工鱼群算法主要模拟了鱼群生活的四大行为：觅食行为、聚群行为、追尾行为和随机行为。其中各种行为的含义如下。

(1) 觅食行为。鱼群中的所有个体(鱼)总体上是趋向于食物运动的。即所有的鱼的最终目的是寻找到水域中营养物质最为丰富的位置。同时假设鱼是通过视觉或味觉来感知水中营养物质浓度的，并且在没有到达食物地点前据此选择移动方向。

(2) 聚群行为。当水域里的鱼的密度不超过一定限度时，鱼群倾向于聚集在一起。表现为鱼群同时觅食或躲避敌害。这种行为在现实中很容易观察到。

(3) 追尾行为。当鱼群之中的某些个体发现食物后，它们附近的鱼都会尾随过来。同时范围逐渐扩大，使附近更多的鱼都意识到食物的地点，并通过追尾行为找到食物。

(4) 随机行为。当鱼群中没有鱼找到营养物质浓度较高的水域的时候，鱼群中的鱼通过在水域中随机游走的方式期望找到食物。

现假设人工鱼群集合为 X，其中第 i 个个体位置可表示为$X_i=(x_i^1,x_i^2,\cdots,x_i^d)$，$i=1,2,\cdots,n$，实际寻优过程中，$x_i^k(k=1,2,\cdots,d)$为寻优变量，$d$ 为寻优变量数量，n 为人工鱼群鱼的数量；令对应于鱼 i 的$Y_i=f(X_i)$，f 为寻优函数，Y 为目标函数值；人工鱼 i 和 j 之间的距离表示为其欧式距离，即$d_{i,j}=\|X_i-X_j\|$；人工鱼群算法为鱼群中的每一个个体设置一个感知范围，记为 v；用 s 表示人工鱼每一次移动的步长；δ 表示某一水域人工鱼的拥挤因子。从而人工鱼群算法可以描述为：

(1) 觅食行为：对于人工鱼 i，在其感知范围内随机寻找一个位置P_j，即从点集合(超球体)$\{P_t|d_{i,t}\leqslant v,i\neq t\}$中随机选择一个元素，判断是否满足条件 $f(P_j)<f(X_i)$(求函数最小值，如果求最大值则相反)。如果满足，则让该人工鱼沿P_j-X_i方向移动一个步长 s，其新位置为X_i'；如果不满足，则重新随机选择位置P_j，并重复判断；反复 T 次后，如果仍不能满足条件，则随机游走一步。

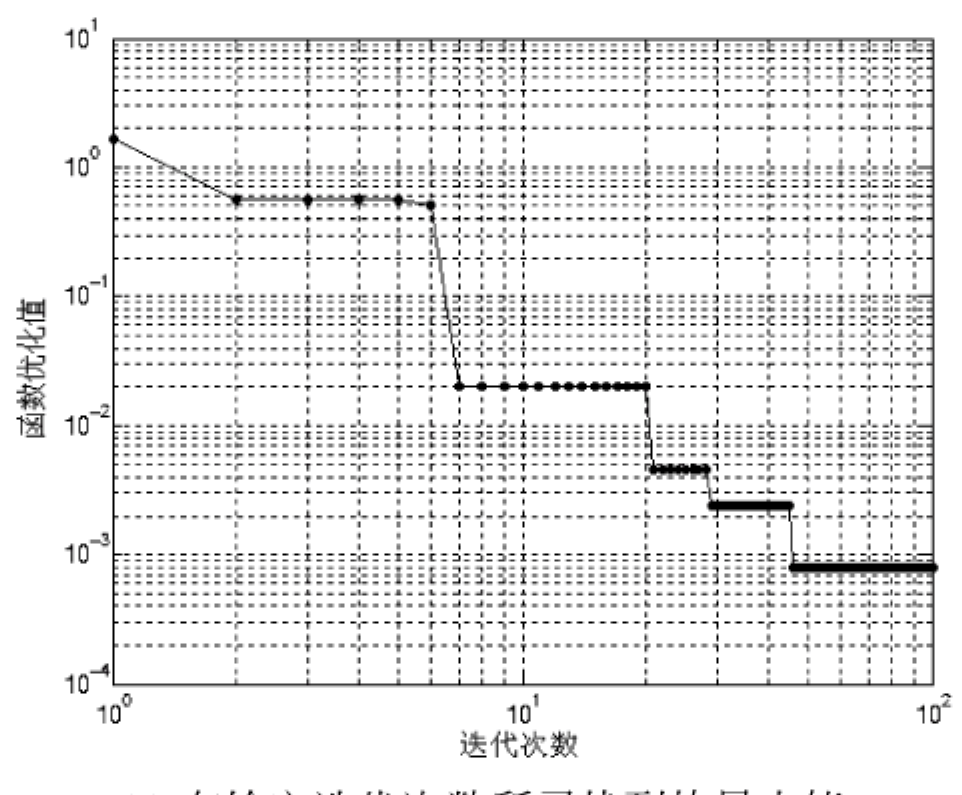
(2) 聚群行为：对于人工鱼 i，在其感知范围内随机寻找人工鱼 j 的集合，即计算人工鱼集合 $B=\{X_t\mid d_{i,t}\leqslant v,t\in X\}$。同时计算 B 的中心位置C，其中 C 的第

k 维为$C_k = \sum x_t^k / \mid B \mid$。如果 $f(C)/\mid B \mid > \delta f(X_i)$，则表明附近的水域的食物足够供给这些鱼，不很拥挤，因而让该人工鱼沿 $C - X_i$ 方向移动一个步长 s；否则执行觅食行为。

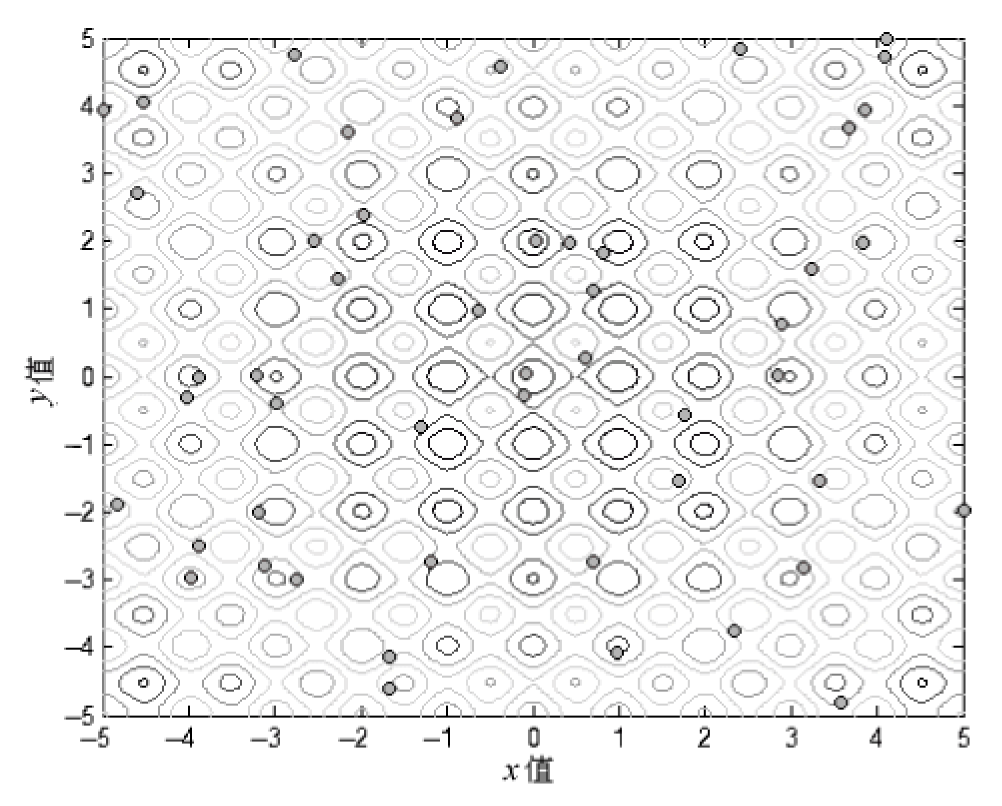
(3) 追尾行为：对于人工鱼 i，在其感知范围内随机寻找具有最大目标函数值 Y_j的人工鱼 j，即寻找人工鱼$\{j \mid Y_j \leqslant \forall Y_t, d_{i,t} \leqslant v, t \in X\}$。如果 $f(X_j)/\mid B \mid > \delta f(X_i)$，则表明$X_j$附近的食物足够充分并且不太拥挤，从而让该人工鱼 j 沿$X_j - X_i$方向移动一个步长 s；否则执行觅食行为。

人工鱼群算法在优化问题上表现出来较强的泛化能力，能够跳出局部最优点，其主要原因在于多个行为之间的协作。觅食行为实际上是人工鱼群算法搜索解空间的主要手段。通过觅食行为，人工鱼群能够从局部到整体非常全面地搜索解空间。聚群行为有助于加强对解空间的全面搜索。当某一位置的人工鱼数量过多时，算法会自组织式地引导人工鱼在其他的位置形成鱼群，从而能够更加有效地跳出局部最优解，提升找到全局最优的可能性。追尾行为则与聚群行为相反，主要致力于局部的搜索。当通过聚群行为大致找到最优点(全局或局部)的位置时，追尾行为能够迅速全面地对这一很小的区域进行彻底的搜索，以很快找到相应位置的最优点。实际上，人工鱼群算法可以看作是一种自组织式的分布式搜索。首先聚群行为将人工鱼按解空间分布，分为若干个小的鱼群；然后追尾行为并行地在这些局部区域内搜索；觅食行为则一方面为前两者提供了实现的手段，另一方面提升了整个解空间的搜索范围。该算法的缺点在于有很多的参数需要确定，特别是在高维解空间的时候，这种问题尤为突出，因而需要进行大量的参数调优工作。

图 8.6 和图 8.7 是使用人工鱼群算法对 Rastrigin 和 Ripple 函数进行优化的结果。可以看出，优化结果较好。最终，人工鱼在解空间的分布较粒子群算法更为均匀。



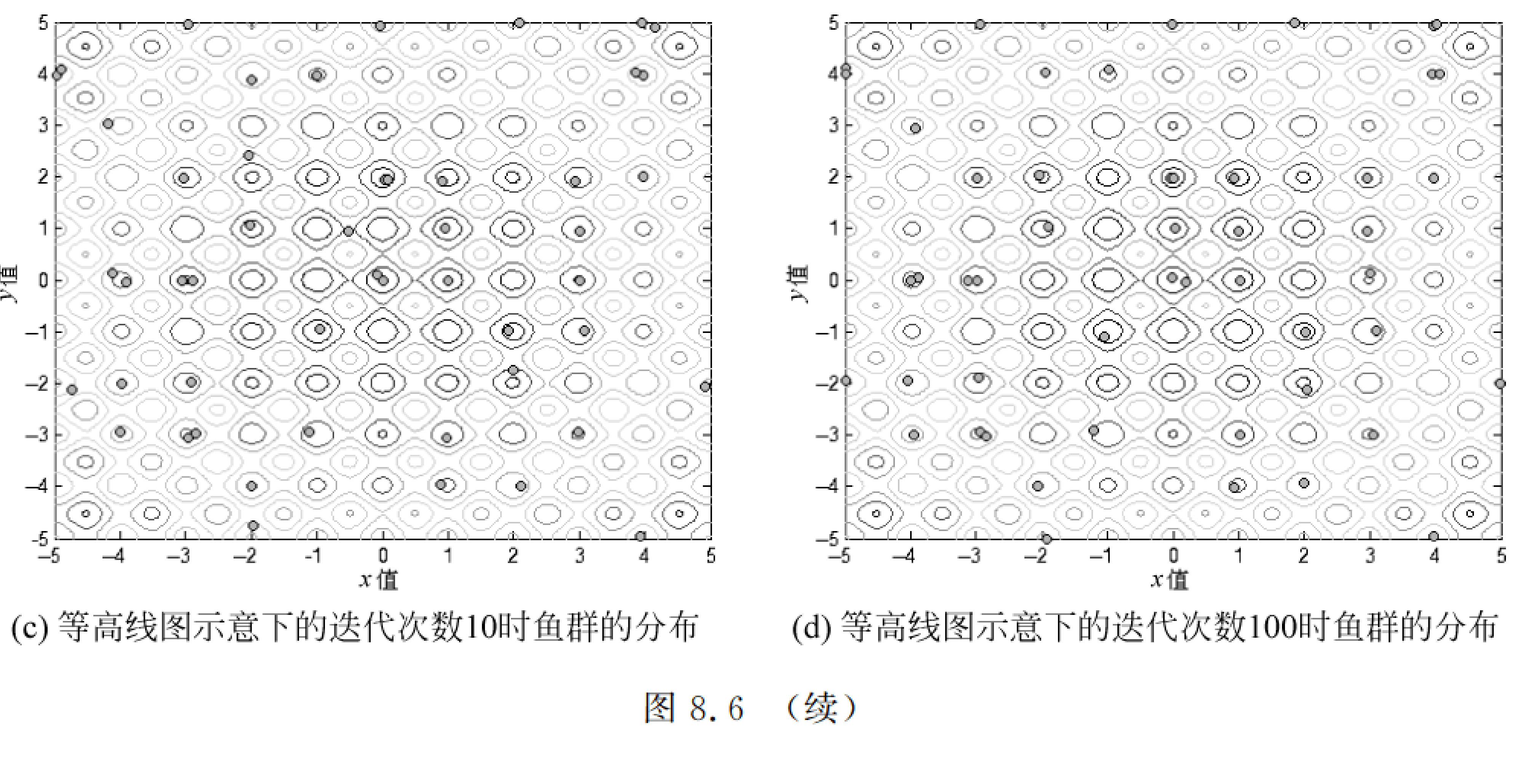

(a) 在给定迭代次数所寻找到的最小值



(b) 等高线图示意下的迭代次数1时鱼群的分布

图 8.6 人工鱼群算法寻找 Rastrigin 函数的最小值

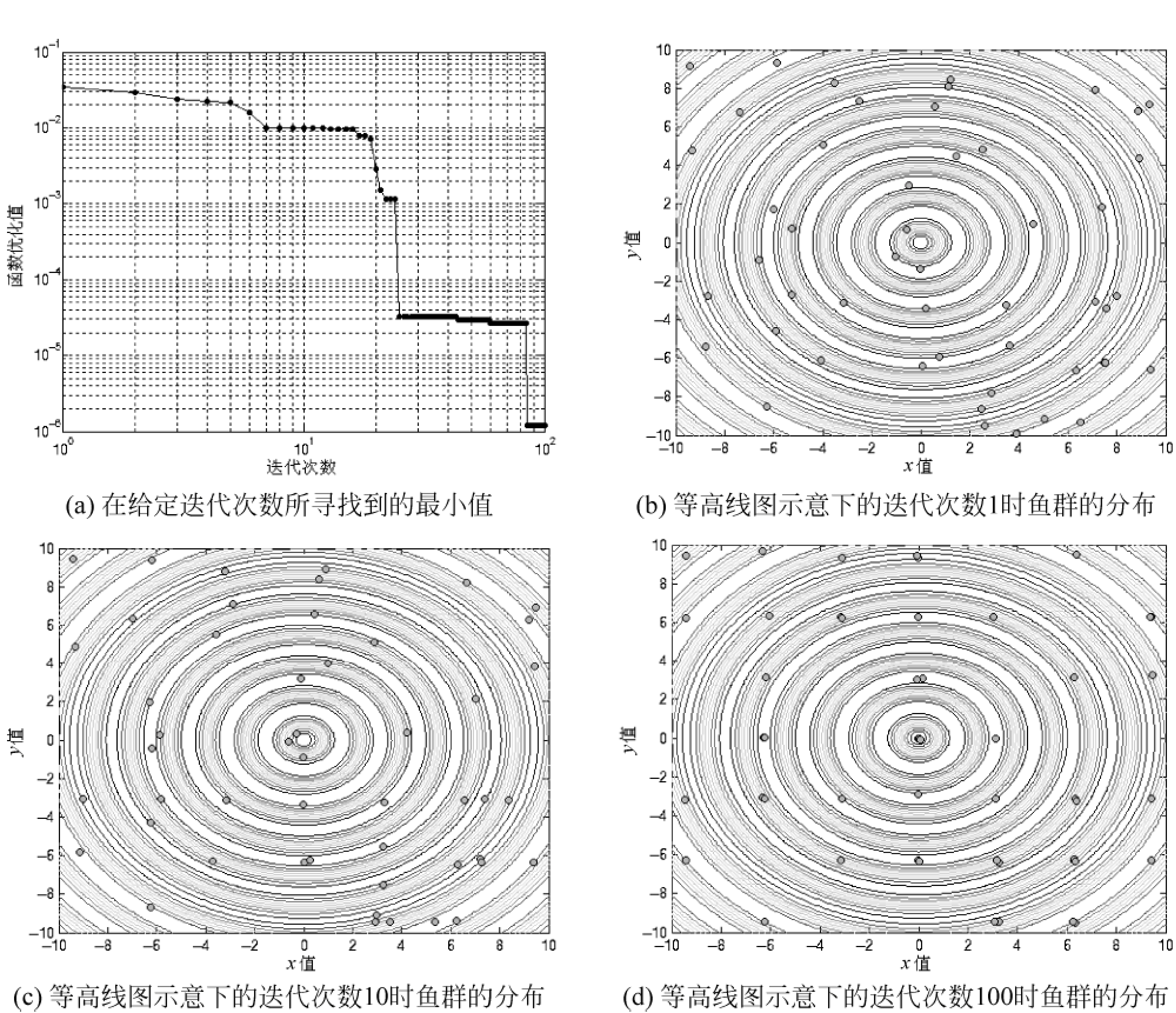
(b，c，d 图中黑色圆点为人工鱼)



(c) 等高线图示意下的迭代次数10时鱼群的分布

(d) 等高线图示意下的迭代次数100时鱼群的分布

图 8.6 （续）



(a) 在给定迭代次数所寻找到的最小值

(b) 等高线图示意下的迭代次数1时鱼群的分布

(c) 等高线图示意下的迭代次数10时鱼群的分布

(d) 等高线图示意下的迭代次数100时鱼群的分布

图 8.7　人工鱼群算法寻找 Ripple 函数的最小值

（b，c，d 图中黑色圆点为人工鱼）

8.4 人工免疫算法

人工免疫算法(Artificial Immune Algorithm，AIA)是模拟生物自然免疫系统响应机制的一种智能算法。该算法由 Farmer 等人于 1986 年首次提出来。但直到 1994 之后，由于 Kephart(1994)的工作才使得该算法开始得到广泛的认知。

科学家和医学工作者很早就发现等到传染病患者痊愈之后，患者就会对这种病有不同程度的免疫力。这种免疫力的产生是机体中的免疫系统的作用。免疫系统具备产生抗体的能力。当病原体(抗原)进入到人体后，体液中的 B 细胞和 T 细胞开始工作。总地来说，T 细胞的作用就是调节其他细胞的活动来对抗抗原，或者是直接对抗原实施攻击。B 细胞则分成两部分，其中一部分成为效应 B 细胞，负责依据 T 细胞呈递的信息针对入侵的病原立即产生抗体，这些抗体能够对病原实施攻击；另外一部分化为长期存活的记忆 B 细胞，它们将存留于机体的血液、淋巴组织中，并在机体中循环，当下一次有相同的病原体入侵时，这些病原体将直接刺激记忆 B 细胞，引起大量的增殖分化，转化为效应 B 细胞，使其迅速产生抗体，消灭抗原，避免机体再次受到同样病原体的攻击。成熟的 T 细胞和 B 细胞分别产生于胸腺和骨髓之中。它们在成熟之后进行克隆增殖、分化表达功能。在免疫系统之中，两种细胞共同作用的同时还相互影响和抑制对方的功能，形成机体内部高度规律的反馈型免疫网络。

总结起来，免疫系统的关键部分在于：

(1) 其中一部分 B 细胞转化为效应 B 细胞，产生抗体；

(2) 另一部分 B 细胞转化为记忆 B 细胞，监控相同病原的入侵，一旦入侵，则迅速大量增殖分化，转化成效应 B 细胞；

(3) 免疫系统在经历病原入侵后存在着超变异，即使机体的免疫能力增强。人工免疫算法即模拟了机体免疫系统中的抗体的产生、抗体与抗原的黏合、克隆、刺激及最终的超变异等过程。

在优化问题求解的过程中，待解决的问题被看作是抗原，问题的解空间为 B 细胞(或抗体)，解的适应度为抗体与抗原的黏合性。忽略 T 细胞对抗抗原和向 B 细胞呈递信息的过程，人工免疫算法主要包括以下几个步骤。

(1) 产生初始的 B 细胞集合B_1。即从待优化问题的解空间随机(或其他方法)生成若干组(N)可行解。

(2) 计算每个 B 细胞和抗原的黏合性。即计算函数在这个解下的函数值。例如求解函数 $Y=f(\mathrm{X})$的最小值，则计算对每个 B 细胞计算 $t=f(\mathrm{B})$，t 值越小则黏合性越强，即适应度越高。

(3) 克隆。从 B 细胞集合B_1中选择黏合性较强的 n 个 B 细胞进行克隆，产生 B 细胞集合B_2。

(4) 变异。对B_2中的 B 细胞进行变异操作，变异的概率随着黏合性的增强而不断变小。变异完成后产生抗体细胞群体B_3。

(5) 选择。从B_1中淘汰黏合性弱的 B 细胞，形成 B 细胞集合B_4。

(6) 更新。从B_3中选择黏合性强的 B 细胞形成集合B_5，将B_5加入到B_4形成新的 B 细胞集。B_5中的 B 细胞每一代都更新，同时淘汰一些相似的 B 细胞。

(7) 重复(2)～(6)直到算法收敛。

从整个算法流程上来看，人工免疫算法和遗传算法具有一定的相似性，但是二者的生物学背景完全不同。遗传算法在给群体中个体提供进化的同时也有相当大的可能性出现明显的退化现象。而人工免疫算法规定了黏合性较高的 B 细胞的变异概率较低，在一定程度上降低了盲目性。

8.5　人本计算

随着社会计算方法得到更深层次的认识，越来越多的人开始构思如何能够更加充分地利用社会大众的力量来完成一些工程量浩大的工作。由人们上网过程中需要输入验证码这一事实而催生的调动全世界范围内的人参加《纽约时报》数字化工作的成功实践就是利用社会计算方法的最为生动的一个事例。

《纽约时报》于 1851 年创刊，距今已有 160 多年的历史，但是能够在网上找到的数字化的内容只有在 1981 年之后，在那之前的 129 年的内容是被扫描成图像而保存在计算机中的。然而这种扫描件的缺点在于一方面占用的存储量巨大，另一方面是无法有效地进行检索。因而，《纽约时报》希望能够将之前的 129 年的内容全部转化为真正意义上的数字存储。但是问题在于这些报纸都非常久远，传统的光学字符识别(Optical Character Recognition, OCR)技术在此没办法保证准确性。当然，如果使用人工录入的话，工作量之巨大是可想而知的。然而最终《纽约时报》神奇地在短短的 24 个月的时间内完成了这一壮举。他们所利用的正是我们所关注的社会计算，在这里又称为“人本计算”。《纽约时报》首先将报纸的扫描件拆分为一段一段的小图片，每个图片上包含 1 个或少量几个单词。之后，他们将这些生成的图片提供给各个网站，让他们将这些图片作为验证字符。这样，全世界各地的人都可以参与解读这些单词。如果达到一定数量的人对某张图片的单词的录入是相同的，那么这就将作为本单词的录入。当所有的单词都被录入后，工作人员只需要将它们重新拼接起来形成完整的文章就可以了。如图 8.8 所示，是一个可能的验证码。

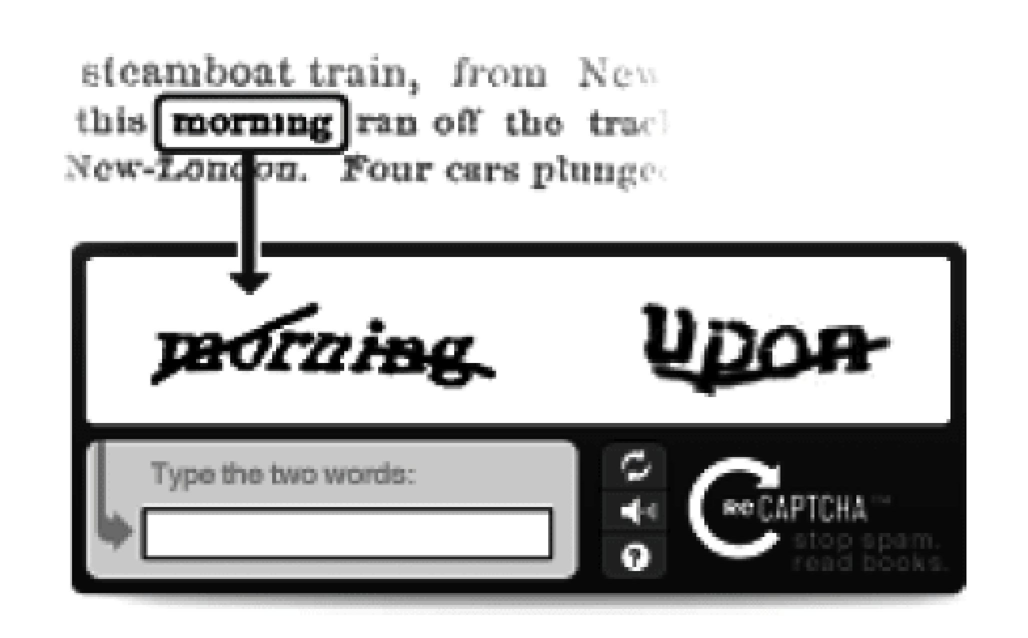


图 8.8　带有 morning upon 单词的验证码

(资料来源：http://www.google.com/recaptcha)

这一方法正是由验证码机制(CAPTCHA)的发明者，卡耐基梅隆大学的著名学者路易斯·凡·安(Luis von Ahn)提出的。他和他的导师发明这一技术的初衷只是防止密码被盗用或是发垃圾邮件。这项技术在发明后的短短几年内就得到了迅速的普及。互联网上，每天被使用的校验码的数量高达 2 亿个。现在，几乎所有的网站都会用到验证码技术。当然，事实证明，路易斯的发明也没有仅仅停留在网络安全应用的层面。路易斯将这一技术拓展之后实际上成就了一门新兴的科学—人本计算(Human-Based Computation)或人计算(Human Computation)。它的基本思想是利用互联网络的分布性和协同性，整合社会的力量来完成单个组织或计算机无法完成的任务。这门科学研究人与人、人与计算机之间的协作，希望把二者的优势都能够发挥出来，从而达到群体智能的效果。《纽约时报》正是这样理念的一个成功实践，由此开发的 reCAPTCHA 系统也正被广泛地使用。

除了 reCAPTCHA 系统外，路易斯还推出过一个著名的游戏，名为 ESP Game。设计的思想非常简单。当某一个人进入游戏后，网站会随机给他分配一个玩家。在每一轮游戏开始后，系统会同时向两个人展示同一张图片。玩家有两分钟的时间对这张图片进行标注，例如海、河、树、天空、猫等。如果两个人标注的关键词有相同的就可以得分，得分奖励能够激励玩家的兴趣。当然，在这背后更为重要的是网站可以根据玩家的标注有效地对这张图片进行标记。例如，如果有 10 个玩家对某一张图片标记了“云”，那么可以肯定“云”应该作为这张图片的一个标签。实际上，为图片做标签是一件非常困难的事情。特别是，当需要对海量的图片进行标记时，在人力和物力双方面基本上都是不可能的。ESP Game 则很完美地解决了这一问题，只要能够保证玩家的数量，图片就能被准确高效地标注。

8.6 补充材料：寻找潜艇“天蝎号”

1968 年 5 月，美国潜艇“天蝎号”在从北大西洋完成例行执勤任务后返回新泽西港的途中突然神秘失踪。美国海军不能判断“天蝎号”到底发生了什么事情，只能根据它最后一次发回的无线电通信信号判断出潜艇的大致位置。最后，海军指挥部只得规划出一个半径达 20 海里，深数千英尺的海域进行搜索。这种搜索无异于大海捞针，潜艇有可能躺在这一海域内的任意一处。时有“美国海军特别计划部首席科学家”头衔的海军军官约翰·克拉文(John Craven)想出了一个独特的搜寻方案。

克拉文邀请了一群具有不同背景知识的专业人士，他们之中包括数学家、潜艇专家、海事搜救等各个领域的专家。依据他们的建议和意见，克拉文编写了“天蝎号”失事的各种不同的可能的“剧本”，同时邀请这些专家依据自己的专业知识对于“天蝎号”依照哪个剧本进行发展做出判断和“投注”。据说，为了让这一过程变得有趣，克拉文还为大家准备了威士忌酒作为投注的奖励。

克拉文最后依据专家们的推断和预测结果得到了潜艇位于所需要搜寻的海域各个位置的概率图。整个概率图被划分为很多个小格子。然后，利用贝叶斯理论对这些小格子代表的区域进行有顺序的搜索。

最终，“天蝎号”在失事 5 个月后被发现。令人吃惊的是，人们最后发现，“天蝎号”实际的失事位置与克拉文依据众专家意见计算得到的预测位置相差的距离仅仅为 220 码。

8.7 本章小结

社会化媒体促使人类生活走向更深入、更全面、更广泛、更便捷的交流模式，缩短了人们彼此之间的距离，为社会网络下的群体智慧的形成和发展奠定了基础。本章从群体智慧的有趣故事开始，分别介绍了蚁群算法、粒子群算法、人工鱼群算法、人工免疫算法、人本计算的机理和实例效果，这些内容揭示了群体运动的组织和发展的重要规律，为社会网络下的群体智能的相关研究提供了新颖的思路。

思考题

1. 简述蚁群算法的机理，试思考社会网络下的蚁群算法的应用场景。
2. 简述粒子群算法的机理，并选择相应的仿真环境，实现利用粒子群算法对

Rastrigin 函数进行优化的实例。

3. 简述人工鱼群算法的机理，并选择相应的仿真环境，实现利用人工鱼群算法对 Ripple 函数进行优化的实例。

4. 请读者查找遗传算法的相关资料，分析人工免疫算法与遗传算法的区别，并选择相关实例进行两者对比。

5. 试思考如何利用群体智慧对社会网络下网民的行为及情感进行建模和分析。

参 考 文 献

[1] Batagelj，Mrvar. Pajek-Program for Large Network Analysis. Connections，1998，21(2)：47-57.

[2] Blei，Ng，Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research，2003，3：993-1022.

[3] Peter Boer，Mark Huisman，Tom A. B. Snijders，et al. StOCNET：An Open Software System for the Advanced Statistical Analysis of Social Networks. ICS/SciencePlus，Universiy of Groningen，Groningen，Netherlands，Version 1，2003.

[4] Stephen Borgatti，Martin Everett，Linton Freeman. UCINET 6.0 Version 1.00. Natick：Analytic Technologies，1999.

[5] Steve Borgatti. NetDraw：Graph Visualization Software. Harvard：Analytic Technologies，2002.

[6] Costa. Hub-based Community Finding. arXiv preprint cond-mat/0405022，2004.

[7] Donetti，Munoz. Detecting Network Communities：A New Systematic and Efficient Algorithm. Journal of Statistical Mechanics：Theory and Experiment，2004(10)：P10012.

[8] Dorigo. Optimization，Learning and Natural Algorithms. Ph. D. Thesis，Politecnico di Milano，Italy，1992.

[9] Dabid Easley，Jon Kleinberg. Networks，Crowds，and Markets：Reasoning about a Highly Connected World. Cambridge University Press，2010.

[10] Farmer，Packard，Perelson. The Immune System，Adaptation and Machine Learning. Physica D：Nonlinear Phenomena，1986，22(1)：187-204.

[11] Gregory. An Algorithm to Find Overlapping Community Structure in Networks. Proceeding of Knowledge Discovery in Databases(PKDD). Springer Berlin Heidelberg，2007：91-102.

[12] Guimera，Sales-Pardo，Amaral. Modularity from Fluctuations in Random Graphs and Complex networks. Physical Review E，2004，70(2)：025101.

[13] Hall. An r-Dimensional Quadratic Placemen Algorithm. Management Science，1970，17(3)：219-229.

[14] Hofmann. Probabilistic Latent Semantic Indexing. Proceeding of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，ACM，1999：50-57.

[15] Kennedy，Eberhart. Particle Swarm Optimization. Proceedings of IEEE IntConf on Neural Networks，1995，4(2)：1942-1948.

[16] Kephart. A Biologically Inspired Immune System for Computers. Proc Artificial Life Ⅳ：The Fourth Int Workshop Synthesis and Simulation of Living Systems，MIT Press，1994：

130-139.
[17] Kermack，McKendrick. Contributions to the Mathematical Theory of Epidemics. Proc Roy Soc. A，1927，115(5)：700-721.
[18] Lancichinetti，Fortunato，Kertész. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics，2009，11(3)：033015.
[19] Liang. An Effective Method of Pruning Support Vector Machine Classifiers. IEEE Trans Neural Networks，2010，21(1)：26-38.
[20] Xun Liang，Hua Chen，Jian Yang. A Method of Detecting and Monitoring Abnormal Internet Information. 美国专利(已授权)，2012，US 8185537.
[21] Liang，Chen，Guo. Pruning Support Vector Machines Without Altering Performances. IEEE Trans Neural Networks，2008，19(10)：1792-1803.
[22] Liang，Ni. Hyperellipsoidal Statistical Classifications in a Reproducing Kernel Hilbert Space. IEEE Trans Neural Networks，2011，22(6)：968-975.
[23] Milgram. The Small-World Problem. Psychology Today，1967，2(1)：60-67.
[24] Newman. Detecting Community Structure in Networks. European Physical Journal (B)，2004，38(2)：321-330.
[25] Newman. Fast Algorithm for Detecting Community Structure in Networks. Physical Review E，2004，69(6)：066133.
[26] Newman，Girvan. Finding and Evaluating Community Structure in Networks. Physical Review E，2004，69(2)：026113.
[27] Pathak，Delong，Banerjee，et al. Social Topic Models for Community Extraction. The 2nd SNA-KDD Workshop'08(SNA-KDD'08)，2008(8).
[28] Radicchi，Castellano，Cecconi，et al. Defining and Identifying Communities in Networks. Proc National Academy of Science，2004，101(9)：2658-2663.
[29] Richardson，Domingos. Mining Knowledge Sharing Sites for Viral Marketing. Proc 8th ACM SIGKDD IntConf Knowledge Discovery & Data Mining，ACM，2002：61-70.
[30] Roweis，Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science，2000，290(5500)：2323-2326.
[31] Schuler. Social Computing. Communications of the ACM，1994，37(1)：28-29.
[32] Shen，Cheng，Cai. Detect Overlapping and Hierarchical Community Structure in Networks. Physica A，2009，388(8)：1706-1712.
[33] Anderson. Customer Satisfaction and Word of Mouth. Journal of Service Research，1998，1(1)：5-17.
[34] James Surowiecki. The Wisdom of Crowds. Random House LLC，2005.
[35] Sznajd-Weron，Sznajd. Opinion Evolution in Closed Community. Int J Modern Physics C，2000，11(6)：1157-1165.
[36] Lei Tang，Huan Liu. 社会计算：社会发现和社会化媒体挖掘. 文益民，闭应洲，译. 北京：机械工业出版社，2012.
[37] Tenenbaum，Silva，Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science，2000，290(12)：2319-2323.

[38] Tyler, Wilkinson, Huberman. E-mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. The Information Society, 2005, 21(2): 143-153.
[39] Vladimir Vapnik. Statistical Learning Theory. Wiley Interscience, 1998.
[40] Wu, Huberman. Finding Communities in Linear Time: a Physics Approach. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 331-338.
[41] 曹润.基于用户联系强度及关注网络的中文微博社区发现研究.中国人民大学硕士学位论文,2013.
[42] 陈华,梁循,阮进.互联网舆情关联分析系统的设计实现.苏州:信息检索与内容安全学术会议论文集,2007:45-49.
[43] 金鑫.基于 MapReduce 的大型数据集聚类方法研究.中国人民大学硕士学位论文,2012.
[44] 李晓磊,邵之江,钱积新.一种基于动物自治体的寻优模式:鱼群算法.系统工程理论与实践,2002,22(11):32-38.
[45] 梁霞.基于支持向量机的股票市场预测建模研究.中国人民大学硕士学位论文,2012.
[46] 梁循,申华,曹润.一种面向微博的全新突发事件发现方法. CN 201210250175.9,2012.
[47] 梁循.数据挖掘算法与应用.北京:北京大学出版社,2006.
[48] 梁循,朱浩然,林航等.基于社会计算的社会化商务模式创新.电子商务,2013(6):20-23.
[49] 林航.基于舆情和支持向量机的股票价格预测.中国人民大学硕士学位论文,2013.
[50] 孟繁荣.社交网络的谣言传播模型研究.南京邮电大学硕士论文,2013.
[51] 倪志豪,梁循,曹润等.一种增量抓取微博信息的方法. CN 201210145247.3,2012.
[52] 宁冉.基于金融信息情感分析的 SVM 股票价格预测.中国人民大学硕士学位论文,2012.
[53] 施晓菁.基于关键词提取的微博用户兴趣模式分析.中国人民大学本科毕业论文,2013.
[54] 施晓菁,梁循,曹润,等.基于兴趣分析的微博博主社区分类方法. CN 201210250181.4,2012.
[55] 王长春,陈超.基于复杂网络的谣言传播模型.系统工程理论与实践,2012,32(1):203-210.
[56] 王超,李楠,李欣丽,等.文本情感倾向性分析用于金融市场波动率的研究.中文信息学报,2009,23(1):95-99.
[57] 王超,梁循.一种互联网新颖词监测方法. CN 200810117821.8,2008.
[58] 徐腾龙.基于复杂网络的微博信息传播模型研究.东华大学硕士学位论文,2013.
[59] 杨源,马云龙,林鸿飞.评论挖掘中产品属性归类问题研究.中文信息学报,2012,26(3):104-108.
[60] 张海燕,余力,梁循.有效评价的多阶段协同推荐研究.北京:全国社会计算学术会议,2012.
[61] 朱浩然.金融领域中文微博情感分析.中国人民大学硕士学位论文,2013.